Journal of Memory and Language 142 (2025) 104603

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Informativity enhances memory robustness against interference in sentence comprehension

Weijie Xu *, Richard Futrell

University of California, Irvine, United States of America

ARTICLE INFO

Dataset link: https://osf.io/e5dsv/

Keywords:
Informativity
Prediction
Strategic memory allocation
Memory interference
Agreement attraction
Sentence processing
Resource-rational

ABSTRACT

Language comprehension has been argued to be expectation-based, with more predictable linguistic units being easier to process. However, as a communicative tool, language is often used to deliver messages that are novel and informative, suggesting the necessity of some cognitive mechanisms handling less predictable but more informative content. This paper proposes strategic memory allocation as one such mechanism. Although less predictable linguistic units require greater processing effort for memory encoding, recognizing the inconsistency between top-down predictions and bottom-up perceptual input may signal the working memory system to prioritize these units, enhancing the robustness of their representation against interference. We examine this hypothesis through the lens of the agreement attraction effect in two self-paced reading experiments. In Experiment 1, we find that less predictable but more informative target nouns exhibit weaker agreement attraction in online reading times, especially with more fine-grained measures of predictability such as the surprisal from large language models. This weaker agreement attraction effect for less predictable target nouns confirms our hypothesis that informative linguistic units are prioritized and receive more robust memory representation. In Experiment 2, however, no modulation of agreement attraction emerges when we manipulate the predictability of distractor nouns, suggesting the need for a more nuanced characterization of how information is structured and operated in memory. Our findings highlight an interplay of memory, predictive processing, and implicit learning. We also discuss the implications of our result for memory efficiency and memory compression. More broadly, by demonstrating that the limited memory resources are dynamically optimized for the relevant processing task, the current study highlights a connection to the resource-rational analysis of human cognition in general.

Introduction

Language use is under the tension between predictability and informativity. On the one hand, comprehenders make predictions about upcoming linguistic units, with less expected units being more difficult to process (DeLong, Urbach, & Kutas, 2005; Kutas & Hillyard, 1980). This observation has inspired an influential line of expectation-based psycholinguistics theories, which seeks to characterize how the difficulty of incremental language processing is shaped by the surprisal of each increment conditioned on its linguistic context (Hale, 2001; Levy, 2008a). While a positive correlation between surprisal and processing difficulty is theoretically and empirically well-established (e.g., Boston, Hale, Kliegl, Patil, and Vasishth 2008, Demberg and Keller 2008, Kuperberg and Jaeger 2016, Shain, Meister, Pimentel, Cotterell, and Levy 2024, Smith and Levy 2013, Wilcox, Gauthier, Hu, Qian, and Levy 2020, Wilcox, Pimentel, Meister, Cotterell, and Levy 2023, Xu, Chon, Liu, and Futrell 2023), it is unlikely to be the full story of how surprisal affects language processing (see also Huang et al., 2024; Huettig & Mani, 2016; Staub, 2024; van Schijndel & Linzen, 2021). From an information-theoretic perspective, surprisal corresponds to information content: linguistic units with higher surprisal in a given context are interpreted as carrying higher information load (Shannon, 1948). Considering that human language functions as a communicative tool for conveying messages that are often novel and informative (Luke & Christianson, 2016), and that the ultimate goal of language comprehension is to extract information, it is crucial to ask whether there exist any cognitive mechanism to handle less predictable but more informative linguistic units in language comprehension.

The role of informativity in sentence comprehension

The role of informativity in comprehension has been previously approached from the perspective of expectation-based processing. Many

* Corresponding author.
E-mail address: weijie.xu@uci.edu (W. Xu).

https://doi.org/10.1016/j.jml.2024.104603
Received 6 March 2024; Received in revised form 11 December 2024; Accepted 16 December 2024
Available online 18 January 2025

studies demonstrate that unexpected information does not always lead to increased processing difficulty. In these studies, predictability and informativity are interpreted in relative terms. That is, information considered implausible and unexpected based on long-term global experience may actually become unsurprising within a specific local context. In support of this idea, it has been found that the comprehender's expectations about the novelty of upcoming information can be reversed, manifested as a decreased processing difficulty associated with globally unpredictable linguistic units (e.g., Arnold, Tanenhaus, Altmann, and Fagnano 2004, Corley, MacGregor, and Donaldson 2007, Hald, Steenbeek-Planting, and Hagoort 2007, Nieuwland and Van Berkum 2006, Rohde, Futrell, and Lucas 2021, Xiang and Kuperberg 2015). For example, consider the sentences below in (1), taken from Xiang and Kuperberg (2015). The verb *celebrated* in a is somewhat unpredictable, since Elizabeth "took the test and failed it," and based on our long-term knowledge this is not something people typically celebrate. However, in the specific local context in b, the phrase *even so* inverses our expectation, and we indeed become more likely to expect the verb *celebrated*.

(1) Global vs. local surprisal (Xiang & Kuperberg, 2015)

a. Elizabeth had a history exam on Monday. She took the test and failed it. She went home and celebrated wildly. [globally surprising]
b. Elizabeth had a history exam on Monday. She took the test and failed it. *Even so*, she went home and celebrated wildly. [locally unsurprising]

However, most of these studies do not directly address the effect of informativity per se. Instead, they focus on how the anticipated level of novelty interacts with the actual level of novelty encountered in a sentence. That is, rather than looking into how sentence processing may benefit from unpredictable but informative linguistic units, these studies mostly investigate how the expectation, or informativity, of a unit can be dynamically changed across different linguistic or non-linguistic environments. In other words, if a globally surprising message becomes unsurprising in a local context, as in (1), the informativity effect has essentially been reduced to a regular predictability effect.

There are only a limited number of expectation-based studies that directly address the concept of informativity per se. Viewing language comprehension as a reverse-engineering process of the speaker's communicative intention (Frank & Goodman, 2012; Goodman & Frank, 2016), some studies propose a comprehension model that integrates the speaker's pressure to deliver novel information. Specifically, they propose a U-shaped effect of informativity, suggesting that linguistic units should carry an intermediate level of informativity that is neither too low nor too high (Rohde et al., 2021). Although this informativity pressure has been documented in many studies from the production side (e.g., Jaeger 2010, Jaeger and Levy 2006, Meister, Pimentel, Wiher, and Cotterell 2023), the empirical support for this effect in comprehension remains limited (Kravtchenko & Demberg, 2022).

Another commonality among these expectation-based studies on informativity is their focus on the immediate impact of surprisal at the moment a linguistic unit is received by the comprehender. However, the real benefit may not show up until later stages of processing. Indeed, implications for this delayed informativity effect can be gleaned from the literature focusing on memory-based sentence processing mechanisms. The idea is that informative linguistic units, after being encoded in memory, can be more readily re-accessed later. One source of evidence for the delayed informativity effect consists in the processing of non-local syntactic dependencies. Consider the sentences in (2), where the antecedent of this non-local dependency *a communist* needs to be re-accessed later from working memory at the retrieval site *banned*:

(2) Semantic complexity facilitates memory retrieval (Experiment 1 in Hofmeister 2011).

a. It was *a communist* who the members of the club banned from ever entering the premises.
b. It was *an alleged Venezuelan communist* who the members of the club banned from ever entering the premises.

Some studies discover that referents with higher semantic complexity, despite their higher difficulty in memory encoding, are actually easier to access at the retrieval site (e.g., Hofmeister 2011, Hofmeister and Vasishth 2014, Karimi, Diaz, and Ferreira 2019, Karimi, Swaab, and Ferreira 2018, Troyer, Hofmeister, and Kutas 2016). For example, in the case of (2), encoding the noun phrase at the beginning of the sentence takes higher processing effort if it features a more complex representation (*a communist* vs. *an alleged Venezuelan communist*). However, at the verb site *banned*, the noun phrase with higher complexity in b can be more easily retrieved, compared to the one in a. Studies of this line, however, often interpret the concept of informativity as semantic or representational complexity, rather than as predictability, which is a common assumption from an information-theoretic perspective. Moreover, it is still under debate whether this facilitated retrieval stems solely from the additional processing effort itself, or also from the enhanced distinctiveness that contrasts the retrieval target with other memory elements (Hofmeister, 2011; Hofmeister & Vasishth, 2014; Karimi, Diaz, & Wittenberg, 2020, 2023).

Hypothesis: Strategic memory allocation based on informativity

The delayed informativity effect outlined above points to a potential interaction between the expectation-based and the memory-based mechanism. Here is our hypothesis. On the one hand, as put forward by the original surprisal theory, less predictable but more informative linguistic units induce greater processing effort when they are first encountered, in that the inconsistency between the top-down prediction and the actual perceptual input needs to be resolved; however, on the other hand, this additional effort might actually help the comprehender construct a memory representation that is more robust against noise and interference in the communicative context. Moreover, we further propose that the rationale behind this hypothesis is a principle that we refer to as **strategic memory allocation**. That is, the limited pool of working memory resources should be strategically allocated during sentence processing to encode, maintain, and retrieve information in memory, prioritizing novel and unexpected information that deviates from top-down predictions.

This strategy of prioritizing unexpected information may arise from a rational process. First and foremost, there is always noise and interference in the communication channel, and there is always the possibility of memory representations being lossy (Brady, Robinson, & Williams, 2024; Gibson, Bergen, & Piantadosi, 2013; Levy, 2008b; Ma, Husain, & Bays, 2014). Despite this seemingly deficient property of memory, through reverse-engineering, language users can in fact use statistical cues to effectively reconstruct the linguistic units whose representation is lost or degraded (Futrell, Gibson, & Levy, 2020; Gibson et al., 2013; Levy, 2008b; Levy, Bicknell, Slattery, & Rayner, 2009; Li & Ettinger, 2023; Ryskin et al., 2021). However, the success rate of this reconstruction is not always the same, and units that are more predictable from the context are more likely to be successfully reconstructed (Futrell, Gibson, & Levy, 2020). When a linguistic unit is lost in memory, comprehenders rely on their prior knowledge to infer what the original sensory input might have been. Among all the possible alternatives, the representation more probable *a priori* is more likely to be selected by the comprehender. In contrast, if the actual representation is not probable in prior knowledge, it is less likely to be selected, thus less likely to be reconstructed. Therefore, to reduce the overall error rate of memory representation, it is rationally

beneficial to avoid memory loss for those unpredictable units in the first place, and this can be done by putting more effort into encoding unpredictable units to enhance the robustness of their representation. This account of strategic memory allocation falls under the resource-rational approach to human cognition in general (Gershman, Horvitz, & Tenenbaum, 2015; Lewis, Howes, & Singh, 2014; Lieder & Griffiths, 2020), a theoretical framework that can be further traced back to the idea of bounded rationality (Simon, 1955). As one branch of the rationalist approach (Anderson, 1990), the resource-rational framework emphasizes both the functional constraints of the external environment and the structural constraints of the internal cognitive system, seeking an optimal, or near-optimal, solution to a computational problem that strikes a balance between them.

Implications for strategic memory allocation can be seen in the resource-rational sentence processing model by Hahn, Futrell, Levy, and Gibson (2022). Their model is trained to make next-word predictions based on a lossy memory context (Futrell, Gibson, & Levy, 2020), where only a limited number of previous words can be preserved. With this memory constraint, they observe that words highly predictable from the context, such as function words, are most likely to be dropped without undermining the model performance in the next-word prediction task. More recently, the idea of strategic memory allocation has been more directly examined in a preliminary analysis by Xu and Futrell (2024b) in the context of dependency locality. In a cross-linguistic corpus study, they find that although there is a general pressure to keep subparts of a syntactic dependency linearly close to each other in order to minimize non-local memory retrieval (e.g., Ferrer-i-Cancho 2004, Futrell, Levy, and Gibson 2020, Gibson 1998, 2000, Hawkins 1994, 2004, Liu 2008), this pressure can be relaxed when the left co-dependent has higher surprisal. They argue that such a relaxed locality pressure may result from strategic memory allocation, where highly surprising co-dependents are prioritized in working memory, making their representations less likely to decay, thus more likely to tolerate longer dependency length.

The allocation of working memory resources has been extensively investigated in other cognitive domains as well. In visual working memory, it has been widely observed that memory resources are not equally distributed, with some information being prioritized to maintain higher representational fidelity (e.g., Bays and Husain 2008, Ma et al. 2014, van den Berg, Shin, Chou, George, and Ma 2012). Importantly, this flexible distribution of memory resources is strongly influenced by the statistical structure of the environment. For example, statistically correlated input data can be stored in a more compressed and abstract form in working memory, without encoding many of the quantitative sensory details (Bates & Jacobs, 2020; Brady, Konkle, & Alvarez, 2009; Brady et al., 2024; Brady & Tenenbaum, 2013). Statistical regularities stored in prior knowledge also play a crucial role. In a delayed-estimation task, Bruning and Lewis-Peacock (2020) show that novel information can be strategically prioritized for memory encoding over familiar information. At a more implementational level, the neural system is adapted to environmental statistics for efficient coding (Atick, 1992; Barlow, 1961; Olshausen & Field, 1996; Rieke, Bodnar, & Bialek, 1995; Simoncelli & Olshausen, 2001; Wiechert, Judkewitz, Riecke, & Friedrich, 2010). Specifically, a group of neurons maximizes the use of its available computing resources by ensuring that every possible combination of neural response levels is equally likely to be used. This efficient coding is achieved by removing redundancies from the original sensory input through a transformation that decorrelates input statistics.

The current study

We examine our hypothesis of strategic memory allocation through the lens of the agreement attraction effect. The phenomenon was first approached from the production side (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991). In English, as illustrated in (3), the head noun *key* of the subject noun phrase, which we refer to as the **target noun**, licenses the number feature of the main verb *was*, forming a subject–verb agreement. However, when there is an intervening **distractor noun** whose number feature mismatches the target noun (*cabinet* vs. *cabinets*), speakers are more likely to make production error as in b, using a main verb that agrees with the distractor noun instead of the target noun. A similar effect is observed from the comprehension side as well (Dillon, Mishler, Sloggett, & Phillips, 2013; Jäger, Mertzen, Van Dyke, & Vasishth, 2020; Pearlmutter, Garnsey, & Bock, 1999; Wagers, Lau, & Phillips, 2009). For comprehenders, the processing difficulty of the ungrammatical main verb in b is reduced when it shares the same number feature with a distractor.

(3) a. The key to the cabinets unsurprisingly was rusty from years of disuse.
b. *The key to the cabinets unsurprisingly were rusty from years of disuse.

The role of predictability, or informativity, in agreement attraction has been mostly examined at the retrieval site, which is the right co-dependent of a non-local dependency (e.g., Lago, Shalom, Sigman, Lau, and Phillips 2015, Parker and Phillips 2017, Wagers et al. 2009). Many studies argue that comprehenders predict the number feature of the verb based on the memory representation of the target noun in the subject. According to these studies, when the actual number feature in the bottom-up perceptual input violates the prediction, cue-based retrieval as a reanalysis process is triggered to resolve the prediction error. This mechanism explains why empirically there is a grammaticality asymmetry in the agreement attraction effect (e.g., Dillon et al. 2013, Hammerly, Staub, and Dillon 2019, Tanner, Nicol, and Brehm 2014, Wagers et al. 2009). That is, in a grammatical sentence, the perceptual input of the verb matches the comprehender's prediction, therefore no retrieval is triggered and no attraction is induced by the distractor. This mechanism also accounts for the observation that the subject–verb agreement exhibits attraction effect more reliably than the reflexives (Dillon et al., 2013): subject–verb agreement is a more predictable dependency than reflexives, making it a more reliable structural cue for retrieval.

Less explored in the literature, however, is how the memory interference can be influenced by the informativity of any linguistic unit outside the retrieval site (Tung & Brennan, 2023). Our hypothesis of strategic memory allocation makes two predictions regarding the informativity of non-retrieval site units. **Our primary prediction** is that the magnitude of the agreement attraction effect should be modulated by the informativity of the target noun, as illustrated in Fig. 1 (top panel). When the target noun conveys novel and unexpected information, we predict that the additional processing effort would help the comprehender to construct a more robust representation of the target noun, reducing the likelihood of interference from the distractor when it needs to be retrieved later. Consequently, we expect a weaker agreement attraction effect for less predictable but more informative target nouns. This primary prediction aligns somewhat with the findings of Hofmeister (2011), which also manipulates the informativity of the retrieval target. However, compared to our study, Hofmeister (2011) focuses more on the retrieval process itself, and the context is held constant across conditions in their stimuli without directly manipulating the interference from the context.

Our secondary prediction is about the informativity of the distractor noun, as illustrated in Fig. 1 (bottom panel). If we assume that the processing of the target noun shares the same limited pool of working memory resources with the distractor, the informativity of the distractor noun should also modulate the agreement attraction effect. Specifically, if more working memory resources are allocated to encoding a novel and unpredictable distractor noun, fewer resources will remain to maintain the representation of the target noun (Brady et al., 2009; Cowan, Rouder, Blume, & Saults, 2012; Thalmann, Souza, & Oberauer, 2019), thereby increasing the likelihood of interference

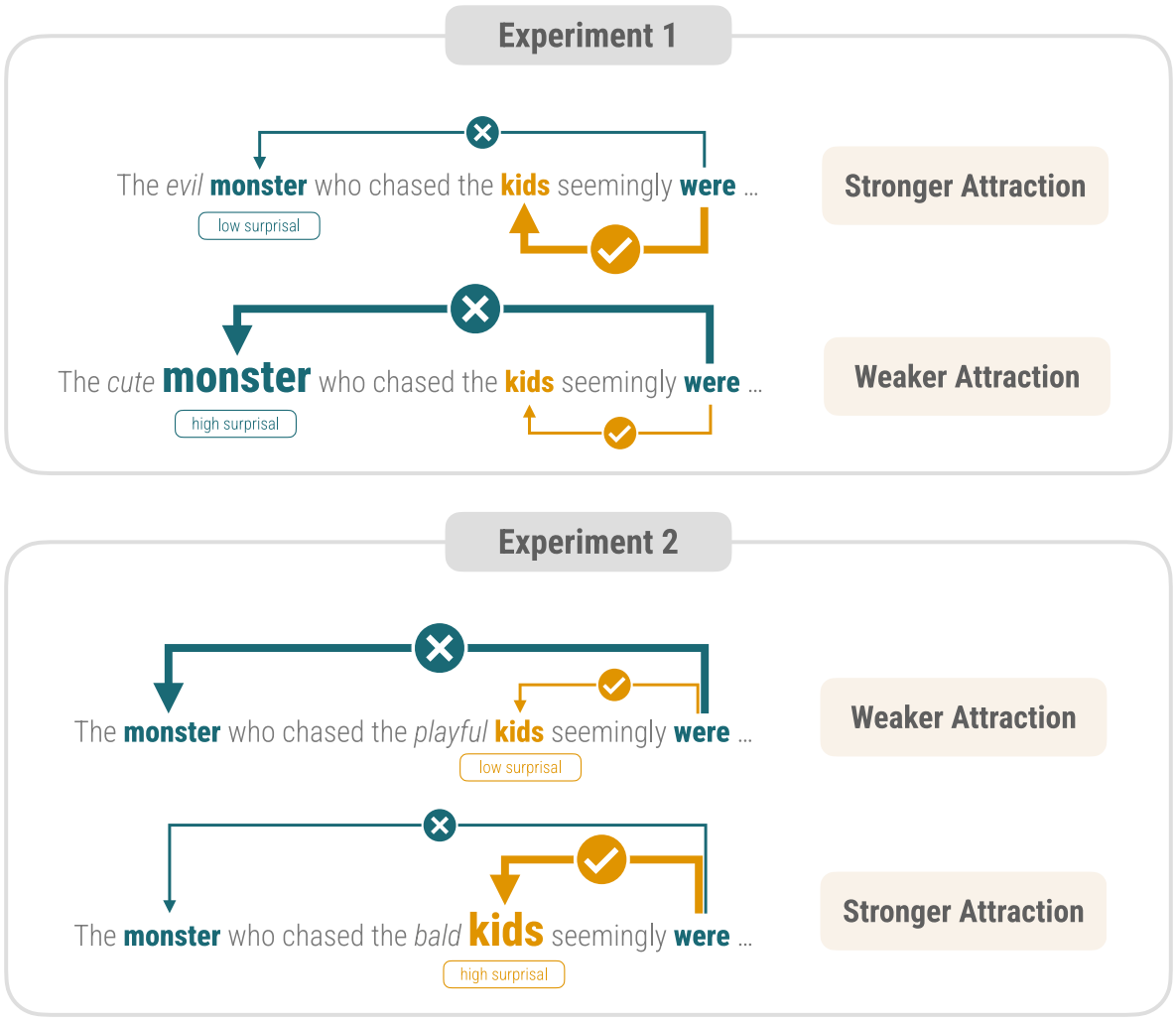

Fig. 1. Strategic memory allocation in agreement attraction. Predictions: target noun of higher surprisal elicits weaker agreement attraction (Experiment 1); distractor noun of higher surprisal elicits stronger agreement attraction (Experiment 2).

from the distractor. As a result, in this secondary prediction, we expect the informativity effect for the distractor noun to operate in the opposite direction, with more informative distractors inducing stronger agreement attraction.

We report two self-paced reading experiments to examine the two predictions above. Experiment 1 aims to test our primary prediction, which posits that the subject–verb agreement with less predictable but more informative target noun should be less susceptible to the interference from the distractor, leading to a weaker agreement attraction effect in comprehension. Experiment 2 aims to test the secondary prediction, which posits that the more informative distractor noun should induce a stronger agreement attraction effect. To preview the results, we find that the informativity of the target noun approximated as the information-theoretic surprisal generated from GPT-2 language model reliably modulates the magnitude of the agreement attraction effect. Specifically, the more surprising the target noun, the weaker the agreement attraction (Experiment 1). However, no reliable modulation of the agreement attraction effect emerges when we manipulate the informativity of the distractor noun, raising the question about how linguistic elements are structured and operated in memory to compete for cognitive resources (Experiment 2).

Experiment 1

This first experiment aims to examine our primary prediction, which posits that less predictable but more informative target nouns yield more robust memory representation, and thus exhibit weaker agreement attraction effect at the retrieval site. In a 2 × 2 × 2 within-subject design, we manipulate the predictability, or surprisal, of the target noun, the grammaticality of the subject–verb agreement (that is, the number feature of the main verb), and the number feature of the distractor noun. Surprisal, as a concept from information theory, corresponds to the level of predictability, with higher surprisal indicating lower predictability (Shannon, 1948). Considering that the agreement attraction effect is predominantly observed with singular target nouns (Bock & Miller, 1991; Pearlmutter et al., 1999; Wagers et al., 2009), the number feature of the target noun is held constant as singular in the current study across all conditions.

Following previous studies (Dillon et al., 2013; Wagers et al., 2009), the agreement attraction effect is instantiated as a `Grammaticality × Distractor` two-way interaction. That is, first, for sentences with a singular subject target noun, there should be an increased processing difficulty if the main verb is in the ungrammatical plural form, compared to its grammatical singular counterpart. However, this grammaticality effect is reduced when there is an intervening plural distractor noun that matches the number feature of the ungrammatical main verb. On top of this baseline agreement attraction effect, we further predict an additional interaction between agreement attraction and target noun surprisal, based on our hypothesis of strategic memory allocation. That is, if novel information indeed enhances the robustness of the target noun's memory representation against the

interference from the distractor noun, we expect a weaker agreement attraction effect associated with more surprising target nouns, as illustrated in Fig. 1 (top panel). This should result in a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction.

Method

Participants

250 English native speakers living in the U.S. were recruited via Prolific and were paid $4.5 at the rate of around $11 per hour for taking the experiment (median = 24.5 min, SD = 14.3 min).

Materials

Table 1 presents a set of sample stimuli.[1] As aforementioned, we manipulated three factors of the subject–verb dependency: (1) predictability, or surprisal, of the target noun; (2) the number feature of the distractor noun; and (3) the grammaticality of the main verb. Each manipulation has two conditions, resulting in eight conditions in total. The manipulation of the target noun surprisal follows the design of the Experiment 3 in Hofmeister (2011): the subject noun phrase that hosts the target noun in the matrix clause is in the form of *Det Adj N*, where the adjective is manipulated into a typical and an atypical condition (for example, *the evil monster* vs. *the cute monster*). The distractor noun in the subject relative clause does not contain any modifier.

The binary manipulation on the target noun gives us the first measure of surprisal, which is a binary categorization of the surprisal of the target noun (high surprisal vs. low surprisal) as in Table 1. We refer to this measure as *Binary Surprisal* in the rest of this article. We also obtained a second measure of surprisal, which is a gradient one, formalized as the negative of the log probability of a word w_t given its preceding context $w_{<t}$:

$$S_t \equiv -\log p\left(w_t \mid w_{<t}\right). \tag{1}$$

We retrieved the estimation of log probability for each target noun from the GPT-2 small (Radford et al., 2019). We refer to this second measure as *GPT-2 Surprisal* in the rest of this article.[2] Transformer-based large language models such as the GPT family provide the state-of-the-art probabilistic measures for next-word prediction, and have been increasingly used in psycholinguistics (Goodkind & Bicknell, 2018; Hao, Mendelsohn, Sterneck, Martinez, & Frank, 2020; Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023; Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Schrimpf et al., 2021; Shain et al., 2024; Wilcox et al., 2023; Xu et al., 2023). Moreover, some recent studies observe that surprisals generated from earlier and smaller models such as GPT-2 correlate better with human behaviors than those from more recent and larger models that are trained on much larger amount of data (Oh & Schuler, 2023; Oh et al., 2024).

There are 32 target items and 96 filler items randomly distributed in the experiment session. Each target item has eight conditions as aforementioned.[3] The target items are latin-square distributed, such that each participant only reads one of the eight conditions for each item and is exposed to all the eight conditions throughout the experiment session. The critical region in the target items is the one that contains the main verb (e.g., *was/were gone*) and the spill-over region immediately follows the critical region (e.g., *before*). Five of the 96 filler items are designed to be infelicitous and serve as attention check.

Procedure

Participants recruited via Prolific were directed to PCIbex (Zehr & Schwarz, 2018) to take the experiment. After being presented with a legal notice and completing a questionnaire on their language background, participants were given the instructions of the experiment. Several questions were asked during this process to help participants better understand the instructions. Six practice trials were given before the main experiment session. In each target trial, participants first read an experimental sentence in the moving-window self-paced reading (SPR) paradigm, and then rated the acceptability for the sentence they read on a 0–100 slider scale, with 100 being the most acceptable. The experimental sentence was not displayed on the screen when participants made acceptability ratings.

Data exclusion

The exclusion of participants consists of two steps. The first step is based on the acceptability judgment rates on filler items. We calculated the average of acceptability judgment rates of infelicitous filler items for each participant, and excluded participants whose average rating over infelicitous fillers is three standard deviations beyond the mean across all the participants (8 participants removed). The second step is based on the online reading times, where participants were excluded if their average reading time per SPR region during the experiment session is three standard deviations beyond the mean among all the participants (5 participants further removed). The trial-level data cleaning for reading time responses also follows two steps. In the first step, we excluded reading time responses that are faster than 100 ms and slower than 4000 ms; in the second step, we further removed reading time data that are beyond three standard deviations from the mean per region and per condition (a total of around 2% reading time data points removed in the critical and the spillover region). For acceptability judgments, we excluded responses that are three standard deviations from the mean per condition (around 0.1% acceptability data points removed).[4]

Data analysis

We ran Bayesian multilevel linear models with `brms` package (Bürkner, 2017) in R on reading times and acceptability judgments. For acceptability judgments, we used beta regression, which is appropriate for data with an upper and lower bound such as a scale of proportion or percentage. The dependent variable in beta regression is assumed to follow a beta distribution, which has two parameters that can be interpreted as the *mean* μ and the *precision* ϕ. The precision is similar to the idea of variance, in the sense that higher precision can be interpreted as more consistent outcomes closer to the mean. Beta regression thus has two components corresponding to the two parameters of a beta distribution. The support of beta distribution is an open interval $(0, 1)$ that does not include 0 and 1. Therefore, following (Smithson & Verkuilen, 2006), acceptability judgments as the dependent variable Y were transformed from its original space $[0, 100]$ to $(0, 1)$ given the

[1] The SPR regions are phrase-by-phrase, such that each region is more or less a legitimate constituent of the sentence. We hope this phrase-by-phrase segmentation can create a more naturalistic reading setup than the word-by-word SPR.

[2] We originally used the GPT-3 base (`text-davinci-001`; Brown et al., 2020) to generate surprisal measures and to develop our stimuli. We later switched to GPT-2 following the reviewers' suggestion, and also for the reproducibility of our result since GPT-3 is no longer accessible from OpenAI since January 2024. In fact, the critical `Grammaticality` × `Distractor` × `Surprisal` three-way interaction has a stronger effect with GPT-2 surprisals than GPT-3, consistent with some recent findings that GPT-2 surprisals better align with human judgments and behaviors (Oh & Schuler, 2023; Oh, Yue, & Schuler, 2024).

[3] In the original development of stimuli, the intended manipulation of binary surprisal (low vs. high) for Item 5, 7, 13, 14, 15, 25, and 26 was later found to be inconsistent with the surprisal estimates from GPT-2. In the analysis using binary surprisal, we corrected this discrepancy by adjusting the labels for these items to align with the GPT-2 surprisal values.

[4] Although this data exclusion process was not preregistered, we performed some sanity checks to ensure that our data exclusion is reasonable and not biased (see details in Appendix "Data Exclusion").

Table 1
Experiment 1 sample stimuli. In a 2 × 2 × 2 design, we manipulated the surprisal of the target noun, the number feature of the distractor, and the grammaticality of the main verb. Slashes indicate phrase-by-phrase SPR regions. The critical and the spillover region for data analysis are underlined. The critical region contains the main verb; the spillover region goes immediately after the critical region.

Low Surprisal Target Noun	
1. *Grammatical, not distracted*	The evil monster/ who/ chased/ the kid/ seemingly/ was gone/ before/ the sunset.
2. *Grammatical, distracted*	The evil monster/ who/ chased/ the kids/ seemingly/ was gone/ before/ the sunset.
3. *Ungrammatical, not distracted*	The evil monster/ who/ chased/ the kid/ seemingly/ were gone/ before/ the sunset.
4. *Ungrammatical, distracted*	The evil monster/ who/ chased/ the kids/ seemingly/ were gone/ before/ the sunset.
High Surprisal Target Noun	
5. *Grammatical, not distracted*	The cute monster/ who/ chased/ the kid/ seemingly/ was gone/ before/ the sunset.
6. *Grammatical, distracted*	The cute monster/ who/ chased/ the kids/ seemingly/ was gone/ before/ the sunset.
7. *Ungrammatical, not distracted*	The cute monster/ who/ chased/ the kid/ seemingly/ were gone/ before/ the sunset.
8. *Ungrammatical, distracted*	The cute monster/ who/ chased/ the kids/ seemingly/ were gone/ before/ the sunset.

following formula:

$$Y' = \frac{(Y/100)(N-1)+0.5}{N}, \quad (2)$$

where N is the sample size.[5] The same linear predictors were used for both the *mean* μ and the *precision* ϕ in our analysis.

For reading time data, we used the standard linear regression that assumes its outcomes to follow Gaussian distribution. We assume that reading times are log-normally distributed. Therefore, the reading time responses were log-transformed before statistical analyses, a common practice in psycholinguistics literature (Burchill & Jaeger, 2024). We analyzed two SPR regions: the critical region where the subject noun phrase is retrieved and where the subject–verb agreement is processed, and the spillover region that immediately follows the critical region.

We report two variants of analysis, corresponding to the two surprisal measures mentioned above, namely the *Binary Surprisal* and the *GPT-2 Surprisal*. In the main analysis, all the categorical variables are effect coded (`Grammaticality`: grammatical 1 vs. ungrammatical −1; `Distractor`: plural 1 vs. singular −1; `Surprisal`: high 1 vs. low −1). The continuous variable is z-scaled. For both measures of surprisal, we included a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction as well as all the corresponding lower-order terms as the predictors in statistical models. For both acceptability judgments and reading times, we included by-item and by-participant random effects with the maximal random structure (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). Following the principles of Bayesian statistics, a strict threshold is not assumed on the posterior probability to *reject* the null hypothesis, and the effect is considered reliable if almost all the probability mass of the posterior distribution for the effect estimate is at one side of zero (Schad, Betancourt, & Vasishth, 2021).[6] See more details of model specification in Appendix "Statistical Models".

[5] Another way to handle value 0 and 1 in the dependent variable is zero-inflated beta regression. It is a mixture of two processes, one being a beta regression to generate outcomes within $(0,1)$, and the other being a logistic regression to generate binary outcomes that are exactly 0 or 1. For the current study, however, it is less motivated to have two separate data-generating processes, and we assume that 0 s and 1 s in the acceptability judgment rates are simply values that are extremely close to the end points of the scale, following the same generative process as other outcomes that are not 0 or 1. Therefore, instead of using zero-inflated beta regression, we applied the regular beta regression after re-scaling the dependent variable to avoid 0 s and 1 s.

[6] Despite this gradient interpretation of posterior probability in Bayesian statistics, when describing the statistical results, we did adopt thresholds in a seemingly frequentist fashion to decide whether to interpret the effect as reliable ($P(|\beta| > 0) > 0.95$), preliminary ($P(|\beta| > 0) > 0.9$), or unreliable ($P(|\beta| > 0) < 0.9$). However, these thresholds are merely heuristic for detecting which effects are likely to exist, and should not be considered strict thresholds to reject the null hypothesis.

Result: Acceptability judgments

Fig. 2 shows the acceptability judgment rates on target items. We focus on the interpretation of the statistical result on the mean μ component of the Bayesian beta regression models, which is summarized in Table 2. The same predictors were used for the precision ϕ component in the model (see the full result with the precision ϕ component in Appendix "Statistical Results on Acceptability Judgments"). The result is qualitatively the same between binary surprisal and GPT-2 surprisal. The target nouns are always singular in all conditions. Therefore, plural distractors should elicit more interference of number feature than singular distractors. There is reliable evidence for all the three main effects: for `Grammaticality`, grammatical sentences are rated more acceptable than ungrammatical ones; for `Distractor`, sentences with the plural distractor are rated more acceptable; for `Surprisal`, sentences with less surprising target noun are rated more acceptable. There is also reliable evidence for a `Grammaticality` × `Distractor` interaction, indicating an agreement attraction effect where the grammaticality effect is reduced when the distractor bears the same number feature as the target noun. The data also support a `Grammaticality` × `Surprisal` interaction, whereby the grammaticality effect is reduced when the target noun is of higher surprisal. There is also a `Distractor` × `Surprisal` interaction, whereby the distractor effect is reduced when the target noun is of higher surprisal. No evidence is found for a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction.

Result: Reading times

Fig. 3 shows raw reading times in the critical and the spillover region. Panel A shows the result with the binary categorization of target noun surprisal. Panel B shows the result with the GPT-2 surprisal of the target noun. As a reminder, plural distractors should elicit more interference of number feature than singular distractors, since the target nouns are in singular form in all conditions. The result of Bayesian statistical models on log-transformed reading times is summarized in Table 3, and the posterior distributions are visualized in Fig. 4.

Binary categorization of surprisal

Critical region. There is a `Grammaticality` main effect, whereby grammatical sentences are read faster than ungrammatical ones. There is also a `Distractor` main effect, whereby plural distractors induce shorter reading times on average than singular ones at the retrieval site. No `Surprisal` main effect is observed. The data support a `Grammaticality` × `Distractor` interaction, indicative of the classic agreement attraction effect where the `Grammaticality` effect is reliably reduced with plural distractors. There is little evidence for either a `Grammaticality` × `Surprisal` or a `Distractor` × `Surprisal` two-way interaction. For the critical `Grammaticality` × `Distractor` × `Surprisal` three-way interaction, the effect is numerically in the expected direction, where the agreement attraction

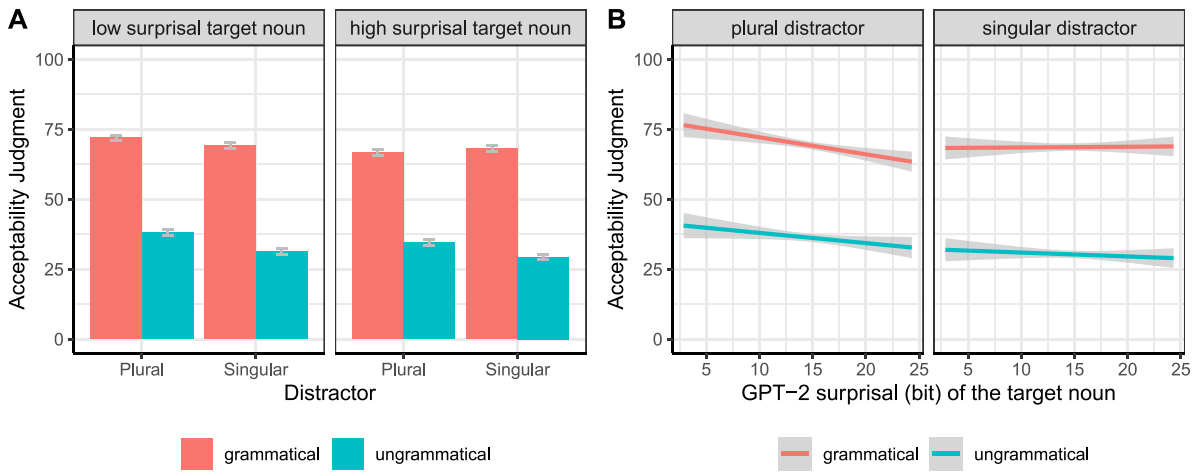

Fig. 2. Experiment 1 acceptability judgment rates. Target nouns are singular in all conditions. *Panel A* corresponds to the binary surprisal (low vs. high) of the target noun; error bars represent standard errors of the mean. *Panel B* corresponds to the GPT-2 surprisal of the target noun.

Table 2
Experiment 1 statistical result of the mean μ component in the Bayesian beta regression on acceptability judgment rates.

Mean μ	Binary surprisal			GPT-2 surprisal		
	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.903	[0.796, 1.013]	$P(\beta > 0) = 1$	0.901	[0.791, 1.010]	$P(\beta > 0) = 1$
Distractor	0.061	[0.035, 0.086]	$P(\beta > 0) = 1$	0.060	[0.034, 0.087]	$P(\beta > 0) = 1$
Surprisal	−0.068	[−0.106, −0.029]	$P(\beta < 0) = 0.999$	−0.100	[−0.163, −0.031]	$P(\beta < 0) = 0.996$
Gram × Distr	−0.045	[−0.071, −0.019]	$P(\beta < 0) = 0.999$	−0.045	[−0.072, −0.018]	$P(\beta < 0) = 0.999$
Gram × Surp	−0.027	[−0.052, −0.003]	$P(\beta < 0) = 0.985$	−0.044	[−0.079, −0.009]	$P(\beta < 0) = 0.992$
Distr × Surp	−0.034	[−0.059, −0.009]	$P(\beta < 0) = 0.996$	−0.030	[−0.056, −0.004]	$P(\beta < 0) = 0.988$
Gram × Distr × Surp	0.004	[−0.021, 0.028]	$P(\beta > 0) = 0.615$	0.004	[−0.023, 0.031]	$P(\beta > 0) = 0.619$

effect instantiated as a Grammaticality × Distractor interaction is numerically reduced for target nouns of higher surprisal. However, this effect is not statistically reliable in this analysis.

Spillover region. First, although there is a Distractor main effect, this effect is in the opposite direction to the one in the critical region. That is, unlike the effect in the critical region, plural distractors now induce longer, not shorter reading times compared to singular distractors. There is no evidence for either a Grammaticality or a Surprisal main effect. In terms of the Grammaticality × Distractor interaction, the effect is also numerically in the opposite direction to the one in the critical region. This suggests that the classic agreement attraction effect originally observed in the critical region is now reversed in the spillover region, whereby the longer reading time associated with plural distractors is numerically more pronounced in ungrammatical sentences. However, this numerically flipped Grammaticality × Distractor interaction does not seem statistically reliable based on the current experiment data. There is no evidence for either a Grammaticality × Surprisal or a Distractor × Surprisal two-way interaction. In the end, there is evidence for a Grammaticality × Distractor × Surprisal three-way interaction in the spillover region, indicating that the reversed Grammaticality × Distractor interaction is more pronounced for high-surprisal target nouns.

Summary. In the critical region, we first successfully replicated the baseline agreement attraction effect manifested as a Grammaticality × Distractor two-way interaction. We also observed a Grammaticality × Distractor × Surprisal three-way interaction, whereby the agreement attraction effect is numerically reduced with more surprising target nouns. However, this effect is not statistically reliable. In the spillover region, there is no longer a classic agreement attraction effect, and the Grammaticality × Distractor two-way interaction is in fact in the opposite direction to the one in the critical region, possibly driven by a reversed Distractor effect, which we will discuss below in Section "Discussion". To sum up, although the reading time pattern is mostly consistent with our predictions in the critical region, the statistical evidence is weak in the current analysis with binary surprisal.

Surprisal from GPT-2 language model

Critical region. For the three main effects, the result is qualitatively the same as the analysis with the binary categorization of surprisals. That is, the data support a Grammaticality and a Distractor main effect, while there is no evidence for a Surprisal main effect. There is also a reliable Grammaticality × Distractor interaction, suggesting a classic agreement attraction effect that is in line with the result with binary surprisal. For the Distractor × Surprisal interaction, unlike the analysis with binary surprisal, there is clearer evidence for a reduced Distractor effect in general when the target noun is of higher surprisal estimates from GPT-2. The target noun surprisal does not modulate the Grammaticality effect, resulting in an absence of Grammaticality × Surprisal interaction for both surprisal measures. Most importantly, with the surprisal estimates from GPT-2, we indeed observed a Grammaticality × Distractor × Surprisal three-way interaction, an effect that is now more reliable compared to the earlier result with binary surprisal. As predicted, this three-way interaction clearly indicates that the classic agreement attraction instantiated as a Grammaticality × Distractor interaction is reliably reduced when the surprisal of the target noun increases.

Spillover region. Unlike the analysis with binary surprisal, there is now preliminary evidence for a Surprisal main effect with GPT-2 surprisal, indicating that more surprising target nouns elicit shorter reading times at the retrieval site. The rest of the effects are qualitatively the same for both surprisal measures. As in the analysis with binary surprisal, the result with GPT-2 surprisal shows a Distractor main effect that is in the opposite direction to the one in the critical region, suggesting that plural distractors induce longer reading times

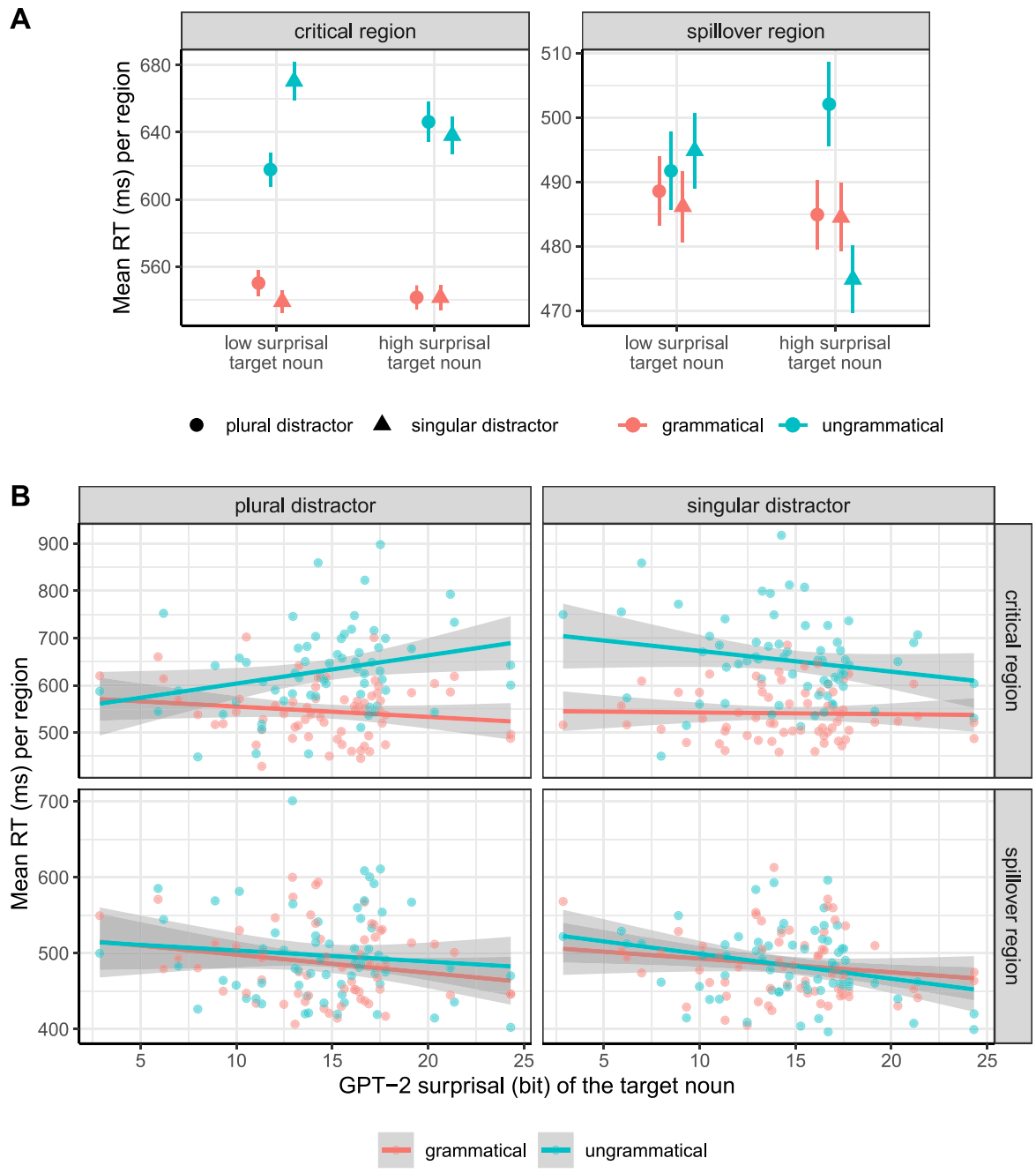

Fig. 3. Experiment 1 reading times in the critical and the spillover region. Target nouns are singular in all conditions. *Panel A* corresponds to the binary surprisal (low vs. high) of the target noun; error bars represent standard errors of the mean. *Panel B* corresponds to the GPT-2 surprisal of the target noun, each point representing the aggregated reading time by item per condition.

than singular ones in the spillover region. The `Grammaticality` × `Distractor` two-way interaction, although not statistically reliable, is numerically in an unexpected direction, too, driven by the reversed `Distractor` effect. There is no evidence for the other two two-way interactions. In the end, although there is still some preliminary evidence for a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction in an unexpected direction, the effect is less prominent compared to the analysis earlier with binary surprisal.

Summary. With GPT-2 surprisal, the reading time result is more clearly consistent with our predictions in the critical region. Again, in the critical region, we first replicated the baseline agreement attraction effect instantiated as a `Grammaticality` × `Distractor` two-way interaction. Most importantly, the data now support a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction, an effect that is more statistically reliable than in the earlier analysis with binary surprisal. This three-way interaction indicates that target nouns of higher surprisal exhibit weaker agreement attraction effect in the critical region. In terms of the spillover region, however, the pattern is still complicated, in the sense that the `Grammaticality` × `Distractor` two-way interaction is numerically in an unexpected direction driven by the reversed `Distractor` effect.

Attraction effect by grammaticality

Given the `Grammaticality` × `Distractor` × `Surprisal` three-way interaction we observed in the critical region in the main

Table 3
Experiment 1 statistical results on log-transformed reading times.

Binary surprisal	Critical region			Spillover region		
	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	−0.064	[−0.074, −0.053]	$P(\beta < 0) = 1$	−0.001	[−0.010, 0.007]	$P(\beta < 0) = 0.613$
Distractor	−0.006	[−0.014, 0.002]	$P(\beta < 0) = 0.936$	0.006	[−0.001, 0.012]	$P(\beta > 0) = 0.959$
Surprisal	−0.003	[−0.011, 0.006]	$P(\beta < 0) = 0.752$	−0.003	[−0.010, 0.003]	$P(\beta < 0) = 0.846$
Gram × Distr	0.009	[0.001, 0.018]	$P(\beta > 0) = 0.982$	−0.004	[−0.010, 0.003]	$P(\beta < 0) = 0.862$
Gram × Surp	0.001	[−0.007, 0.009]	$P(\beta > 0) = 0.596$	0.001	[−0.006, 0.008]	$P(\beta > 0) = 0.584$
Distr × Surp	0.005	[−0.003, 0.014]	$P(\beta > 0) = 0.897$	0.003	[−0.004, 0.011]	$P(\beta > 0) = 0.829$
Gram × Distr × Surp	−0.005	[−0.013, 0.003]	$P(\beta < 0) = 0.896$	−0.006	[−0.013, 0.001]	$P(\beta < 0) = 0.951$
GPT-2 surprisal	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	−0.064	[−0.075, −0.054]	$P(\beta < 0) = 1$	−0.002	[−0.011, 0.006]	$P(\beta < 0) = 0.683$
Distractor	−0.007	[−0.015, 0.002]	$P(\beta < 0) = 0.943$	0.006	[−0.001, 0.012]	$P(\beta > 0) = 0.952$
Surprisal	−0.004	[−0.017, 0.009]	$P(\beta < 0) = 0.75$	−0.007	[−0.017, 0.003]	$P(\beta < 0) = 0.909$
Gram × Distr	0.009	[0.000, 0.018]	$P(\beta > 0) = 0.981$	−0.003	[−0.010, 0.003]	$P(\beta < 0) = 0.852$
Gram × Surp	−0.005	[−0.014, 0.005]	$P(\beta < 0) = 0.849$	0.001	[−0.007, 0.009]	$P(\beta > 0) = 0.583$
Distr × Surp	0.007	[−0.002, 0.018]	$P(\beta > 0) = 0.932$	−0.001	[−0.008, 0.006]	$P(\beta < 0) = 0.637$
Gram × Distr × Surp	−0.011	[−0.020, −0.001]	$P(\beta < 0) = 0.986$	−0.005	[−0.013, 0.003]	$P(\beta < 0) = 0.904$

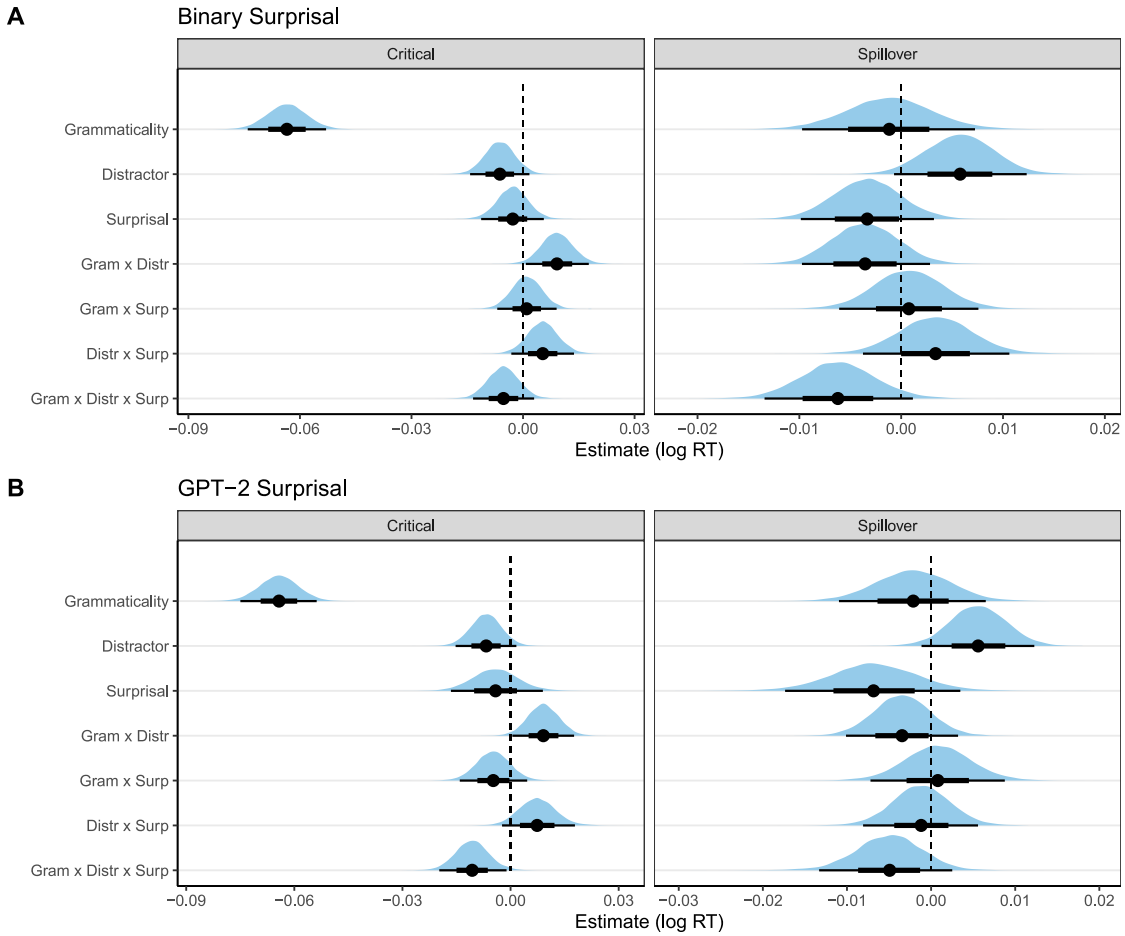

Fig. 4. Experiment 1 posterior distributions with log-transformed reading times. *Panel A*: binary surprisal (low vs. high) of the target noun; *Panel B*: GPT-2 surprisal of the target noun. Black circles represent the mean of posterior estimates. Error bars represent 66% (thick) and 95% (thin) credible intervals.

analysis above, we further looked into whether this effect is mainly driven by ungrammatical or grammatical sentences. That is, in this additional analysis, we looked into the agreement attraction within grammatical and ungrammatical sentences separately, and we examined their interaction with the surprisal of the target noun. Compared to the main analysis above, the agreement attraction effect, instead of being instantiated as a `Grammaticality` × `Distractor` two-way interaction, is represented by the predictor `Distractor` alone but estimated within each `Grammaticality` condition. Therefore, we used contrast coding with nested effects in this analysis, as in Table 4, following Nicenboim, Schad, and Vasishth (2024). The nested coding results in a factor with three levels that contrasts: (1) plural and singular distractors within ungrammatical sentences (`Distractor` in ungrammatical); (2) plural and singular distractors within grammatical sentences (`Distractor` in grammatical); and (3) grammatical and ungrammatical sentences (`Grammaticality` main effect). With this nested coding, the statistical model in this analysis is specified in a way that estimates how the target noun surprisal interacts with each of these three levels. Details of model specification are in Appendix "Statistical Models".

Table 4
Nested contrast coding for the attraction effect within each condition of `Grammaticality`.

Condition	`Distractor` in ungram	`Distractor` in gram	`Grammaticality`
grammatical, plural	0	1/2	1/2
grammatical, singular	0	−1/2	1/2
ungrammatical, plural	1/2	0	−1/2
ungrammatical, singular	−1/2	0	−1/2

Table 5
Experiment 1 attraction effect within each grammaticality condition (nested coding) in the critical region.

	Binary surprisal			GPT-2 surprisal		
	Estimate	95% CrI	Post probability	Estimate	95% CrI	Post probability
Grammaticality	−0.127	[−0.148, −0.106]	$P(\beta < 0) = 1$	−0.129	[−0.150, −0.108]	$P(\beta < 0) = 1$
Surprisal	−0.003	[−0.011, 0.006]	$P(\beta < 0) = 0.75$	−0.004	[−0.017, 0.009]	$P(\beta < 0) = 0.75$
Gram × Surp	0.002	[−0.014, 0.018]	$P(\beta > 0) = 0.597$	−0.010	[−0.028, 0.009]	$P(\beta < 0) = 0.856$
Distractor (ungrammatical)	−0.031	[−0.055, −0.006]	$P(\beta < 0) = 0.992$	−0.032	[−0.057, −0.006]	$P(\beta < 0) = 0.992$
Distractor (grammatical)	0.006	[−0.018, 0.028]	$P(\beta > 0) = 0.689$	0.005	[−0.018, 0.028]	$P(\beta > 0) = 0.667$
Distr × Surp (ungrammatical)	0.021	[−0.003, 0.046]	$P(\beta > 0) = 0.955$	0.036	[0.007, 0.066]	$P(\beta > 0) = 0.991$
Distr × Surp (grammatical)	−0.000	[−0.014, 0.018]	$P(\beta < 0) = 0.507$	−0.007	[−0.033, 0.019]	$P(\beta < 0) = 0.73$

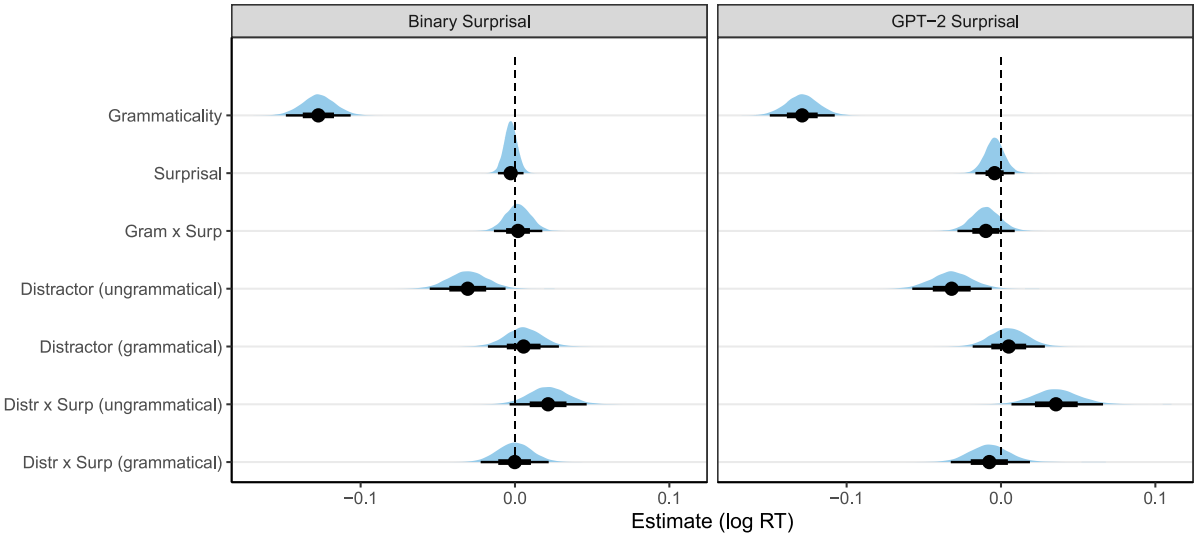

Fig. 5. Experiment 1 posterior distributions for the attraction effect within each grammaticality condition at the critical region. Black circles represent the mean of posterior estimates. Error bars represent 66% (thick) and 95% (thin) credible intervals.

As summarized in Table 5 and visualized in Fig. 5, the statistical results in this analysis are qualitatively the same for both surprisal measures. First, there is a `Grammaticality` main effect, whereby ungrammatical sentences are read more slowly than their grammatical counterparts. There is little evidence either for a `Surprisal` main effect or for a `Grammaticality` × `Surprisal` interaction. For the baseline attraction effect, which is represented by the predictor `Distractor` in this analysis, the data support a reliable attraction effect only within ungrammatical sentences, whereby the ungrammatical main verb is read faster when there is a plural distractor. No baseline attraction effect was observed within grammatical sentences. More importantly, there is reliable evidence for a reduced attraction effect for surprising target nouns within ungrammatical sentences. No modulation of the baseline attraction effect was observed within grammatical sentences.

Discussion

We highlight two critical empirical findings of this experiment. First, we successfully replicated the baseline agreement attraction effect both in acceptability ratings and in the reading time responses, whereby the plural distractor makes sentences with ungrammatical plural main verbs more acceptable in offline judgment and easier to process in the critical reading time region. Second, the most important finding of this first experiment is that target nouns of higher surprisal induce weaker agreement attraction effect in the reading time responses of the critical region. A closer look into the by-item and by-subject effects shows that the effect is not dominantly driven by a small subset of participants or experimental items (See Fig. C.1 in Appendix "Experiment 1 by-item and by-subject effects"). This finding indicates that the number feature of the less predictable but more informative target noun is less susceptible to the interference from the intervening distractor noun, and therefore can be more accurately retrieved at a later point in the sentence. In support of our hypothesis of strategic memory allocation, this finding suggests that less predictable linguistic units, although requiring more cognitive resources to process during encoding, may turn out to yield a more robust memory representation against noise and interference in the context. This target noun surprisal effect on agreement attraction, however, is not observed in acceptability judgment data.

Moreover, both the baseline agreement attraction effect and its interaction with the target noun surprisal are more reliably observed in ungrammatical sentences than in their grammatical counterpart. As pointed out by previous studies, such a grammaticality asymmetry challenges the view that similarity-based interference is mainly caused by feature distortion in the representation of the target noun, in that the originally grammatical sentences should have been perceived ungrammatical if the number feature of the target noun is overwritten by that of the distractor (see Hammerly et al. (2019) for a different view

on the grammaticality asymmetry). A test that directly probes offline interpretations may be needed in the future, in order to see whether the effect we observed is indeed induced by feature overwriting in the representation of the target noun (Dempsey, Christianson, & Tanner, 2022; Patson & Husband, 2016). We also acknowledge that we cannot rule out the possibility that such a grammaticality asymmetry is an artifact of the design of this experiment, where the subject head nouns are always singular. However, it is commonly observed that plural target nouns are rarely attracted by singular distractors, making it even less likely to observe a grammaticality asymmetry (Bock & Miller, 1991; Pearlmutter et al., 1999; Wagers et al., 2009).

It is worth noting that the reading time pattern in the spillover region is complicated by the unexpectedly reversed `Distractor` effect. That is, plural distractors used to yield shorter reading times in the critical region due to agreement attraction, but now yield longer reading times than singular ones in the spillover region. This reversed distractor effect is numerically more pronounced for ungrammaitcal sentences and for high-surprisal target nouns, resulting in a `Grammaticality` × `Distractor` × `Surprisal` three-way interaction effect. This reversed effect seems to point to a trade-off between the critical and the spillover region. One possible explanation is that even though plural distractors can initially elicit agreement attraction in the critical region, comprehenders may later realize their misprocessing and take additional time to reanalyze this misprocessing in the spillover region. Moreover, this reanalysis effect, if it exists, seems to be more pronounced with high surprisal target nouns as evidenced by the three-way interaction. Potentially consistent with our hypothesis of strategic memory allocation, this effect may indicate that high surprisal target nouns, even if they are erroneously processed initially due to the distractor, are still possibly easier for comprehenders to later realize their misprocessing since their representation is more robust. We acknowledge that the current study is not designed to directly test this hypothesis, but it is worth investigating this later-stage processing in future work in greater detail.

Another caveat, as mentioned above, is that high-surprisal target nouns exhibit reduced agreement attraction effect only in online reading time data, but not in offline acceptability judgments. We speculate that this inconsistency may be because online and offline data are measuring different psychological processes. Reading time data, on the one hand, reflect to what extent comprehenders have detected the inconsistency of number feature on the subject–verb dependency and how much effort they have put into addressing this inconsistency, whereas the offline acceptability judgments, on the other, more reflect the ultimate interpretation. There are in fact many decision points from detecting an agreement error to generating the ultimate interpretation (Paape, Avetisyan, Lago, & Vasishth, 2021). For example, the sentence may be judged as bad as it is if an error is detected. Or, an error correction may take place to re-write the number feature, and the sentence may be ultimately judged acceptable even though an agreement error was indeed detected. All these processes are potential confounds that may obscure the pattern of offline data. In order to fully understand the offline data, it is necessary to directly probe the content of the memory representations and to have a more detailed characterization of the linkage between online and offline data in future work.

To sum up, in Experiment 1, we successfully replicated the baseline agreement attraction effect, and mostly importantly, as predicted, we observed that this agreement attraction effect is reduced when the target noun is of higher surprisal. In support of our hypothesis of strategic memory allocation, this finding suggests that linguistic units carrying less predictable but more informative content have more robust memory representation against interference, possibly because they are prioritized for cognitive resources in memory encoding. We also noted two caveats in our result. First, the critical effects are only observed in online reading time data in the critical SPR region. Second, although the critical effects are in the predicted direction in the critical SPR region, the effects seem to be reversed in the spillover region.

Experiment 2

In the second experiment, we examine our secondary prediction: if more memory resources are used to encode the less predictable but more informative distractor noun, fewer resources will be left for the target noun, thus increasing the likelihood of interference from the distractor at the retrieval site. If this is the case, we predict that the informativity of the distractor operates in the opposite direction to that of the target noun, such that more surprising distractor noun should induce a stronger agreement attraction effect, as illustrated in Fig. 1 (bottom panel). As in Experiment 1, the target noun in Experiment 2 is singular in all conditions. However, compared to Experiment 1, the distractor noun in Experiment 2 is always in plural form, which means that there is always high interference from the distractor in terms of the number feature. Therefore, instead of a 2 × 2 × 2 design, Experiment 2 is in a 2 × 2 within-subject manipulation. We make this modification in order to, first, increase the statistical power to better detect the expected effect if it exists. Second, we dropped the conditions with singular distractors while maintaining the manipulation of grammaticality in order to minimize any potential confounding effect from reading too many ungrammatical sentences. With this simplified design, the agreement attraction effect is manifested as a grammaticality illusion, such that ungrammatical sentences will be perceived more grammatical when there is stronger attraction. Therefore, since we expect stronger agreement attraction to be associated with more surprising distractor nouns, we predict that the more surprising plural distractor noun should induce stronger grammaticality illusion, resulting in a `Grammaticality` × `Surprisal` two-way interaction.

Method

Participants

105 English native speakers living in the U.S. were recruited via Prolific and were paid $4.5 at the rate of around $12.6 per hour for taking the experiment (median=21.5 min, SD=12 min).

Materials and procedure

The materials of Experiment 2 are adapted from Experiment 1. In a 2 × 2 within-subject design, we manipulated the grammaticality on the main verb and the surprisal of the distractor noun. A set of sample stimuli is presented in Table 6. We maintain the same number of experimental items as in Experiment 1 (N=32).[7] Compared to Experiment 1, Experiment 2 only includes plural distractors, such that the number feature on the distractor noun always mismatches the one on the target noun, and therefore there is always certain degree of interference from the distractor in terms of number feature. In terms of surprisal, similar to Experiment 1, we manipulated the surprisal of the distractor noun through a pre-nominal adjective, so the distractor NP in the relative clause is in the form of *Det Adj N*. There is no adjective before the target noun in this experiment. Again, we used two measures of surprisal in the subsequent analysis, namely the *Binary Surprisal* and the *GPT-2 Surprisal*. The GPT-2 surprisals of the distractor noun were generated based on the context within the distractor NP, for example $-\log p$(“kids” | “the playful”).[8] There are 64 filler items. As in Experiment 1, five filler items were designed to be infelicitous and served as attention check. The data collection and experiment procedure are identical to Experiment 1.

[7] Before the current Experiment 2, we have run another experiment with the same 2 × 2 within-subject design, but with smaller sample size (60 participants) and with only 16 experimental items. The current Experiment 2 is a replication with more items and participants in order to boost the statistical power. The result did not qualitatively change.

[8] An alternative way to generate the GPT-2 surprisal of the distractor noun is based on the context that includes the entire prefix, for example $-\log p$(“kids” | “The monster who chased the playful”). Since the statistical result does not qualitative differ between these two types of context, we only report the result with the surprisal generated from within the distractor NP.

Table 6
Experiment 2 sample stimuli. In a 2 × 2 design, we manipulated the surprisal of the distractor noun, the number feature of the distractor, and the grammaticality of the main verb. Slashes indicate phrase-by-phrase SPR regions. The critical and the spillover region for data analysis are underlined. The critical region contains the main verb; the spillover region goes immediately after the critical region.

Low Surprisal Distractor Noun	
1. *Grammatical*	The monster/ who/ chased/ the playful kids/ seemingly/ was gone/ before/ the sunset.
2. *Ungrammatical*	The monster/ who/ chased/ the playful kids/ seemingly/ were gone/ before/ the sunset.
High Surprisal Distractor Noun	
3. *Grammatical*	The monster/ who/ chased/ the bald kids/ seemingly/ was gone/ before/ the sunset.
4. *Ungrammatical*	The monster/ who/ chased/ the bald kids/ seemingly/ were gone/ before/ the sunset.

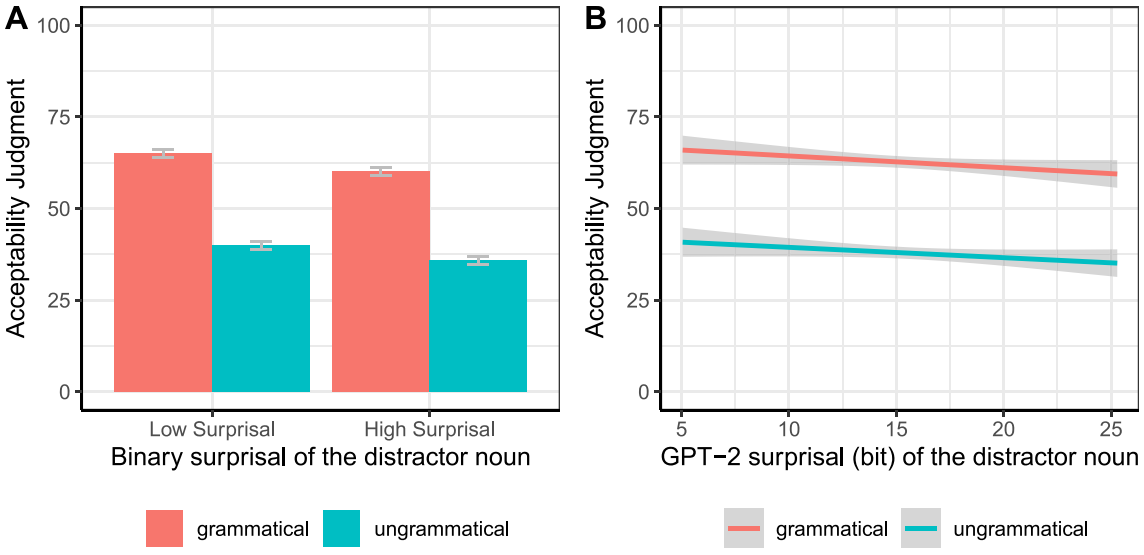

Fig. 6. Experiment 2 acceptability judgment rates. Target nouns are singular in all conditions. Distractors are plural in all conditions. *Panel A* corresponds to the binary surprisal (low vs. high) of the distractor noun; error bars represent standard errors of the mean. *Panel B* corresponds to the GPT-2 surprisal of the distractor noun.

Data exclusion and analysis

As in the previous experiment, the subject-level exclusion is based on both the acceptability judgment rates and reading times. For acceptability judgments, the exclusion consists of two steps. In the first step, participants were removed if their mean acceptability rate for infelicitous fillers is above 50/100 (8 participants removed). Then, in the second step, participants were removed if their mean acceptability rating over infelicitous fillers is 3 standard deviations away from the mean across all the participants (4 participants removed).[9] The trial-level data exclusion follows the same procedure as in Experiment 1. Around 2.3% of the reading time responses were removed in the critical and the spillover region; around 0.07% of the acceptability judgments were removed. The data analysis on acceptability judgments and reading times follows the same method as in Experiment 1. Again, we ran analysis based on both the binary surprisal and the GPT-2 surprisal of the distractor noun. See details of the specification of statistical models in Appendix "Statistical Models".

Result: Acceptability judgments

Fig. 6 shows acceptability judgment rates on the target items. As mentioned above, Experiment 2 is in a simplified design that only includes plural distractors and the target nouns are always singular in all conditions. As in Experiment 1, here we focus on the statistical result on the mean μ component of the Bayesian beta regression models, which is summarized in Table 7 (see the full result with the precision ϕ component in Appendix "Statistical Results on Acceptability Judgments"). The result with binary surprisal does not qualitatively differ from the one with GPT-2 surprisal. There is a `Grammaticality` main effect, whereby grammatical sentences are rated more acceptable than ungrammatical ones. There is also a `Surprisal` main effect, whereby sentences with more surprising distractors noun are rated less acceptable. There is no evidence for a `Grammaticality` × `Surprisal` two-way interaction effect.

Result: Reading times

Fig. 7 shows reading times with the binary surprisal and the GPT-2 surprisal of the distractor noun in the critical and the spillover region. As a reminder, distractors are always plural, and target nouns are always singular across all conditions. The result of Bayesian statistical models on log-transformed reading times is summarized in Table 8, and the posterior distributions are visualized in Fig. 8.

Critical region. First, there is a `Grammaticality` main effect with both surprisal measures, whereby ungrammatical sentences are read more slowly. Second, there is also a distractor `Surprisal` main effect with both surprisal measures, whereby distractors of higher surprisal result in faster reading times at the retrieval site. In the end, we observed preliminary evidence for a `Grammaticality` × `Surprisal` interaction in the analysis with binary surprisal, whereby the `Grammaticality` effect is reduced when the distractors are of higher surprisal. This interaction effect, however, is not observed in the analysis with GPT-2 surprisal.

[9] In Experiment 2, subject-level exclusion based on acceptability ratings involved an additional step compared to Experiment 1. Specifically, before the exclusion based on standard deviation, we first excluded participants whose mean rating over infelicitous fillers was above 50/100. This additional step was necessary because the second step alone in Experiment 2 could not exclude all inattentive participants who assigned disproportionately high ratings (that is, above 50/100) to the infelicitous filler items.

Table 7
Experiment 2 statistical result of the mean μ component in the Bayesian beta regression on acceptability judgment rates..

Mean μ	Binary surprisal			GPT-2 surprisal		
	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.585	[0.448, 0.723]	$P(\beta > 0) = 1$	0.581	[0.438, 0.724]	$P(\beta > 0) = 1$
Surprisal	−0.118	[−0.162, −0.073]	$P(\beta < 0) = 1$	−0.172	[−0.237, −0.107]	$P(\beta < 0) = 1$
Gram × Surp	−0.004	[−0.041, 0.034]	$P(\beta < 0) = 0.573$	−0.026	[−0.077, 0.026]	$P(\beta < 0) = 0.839$

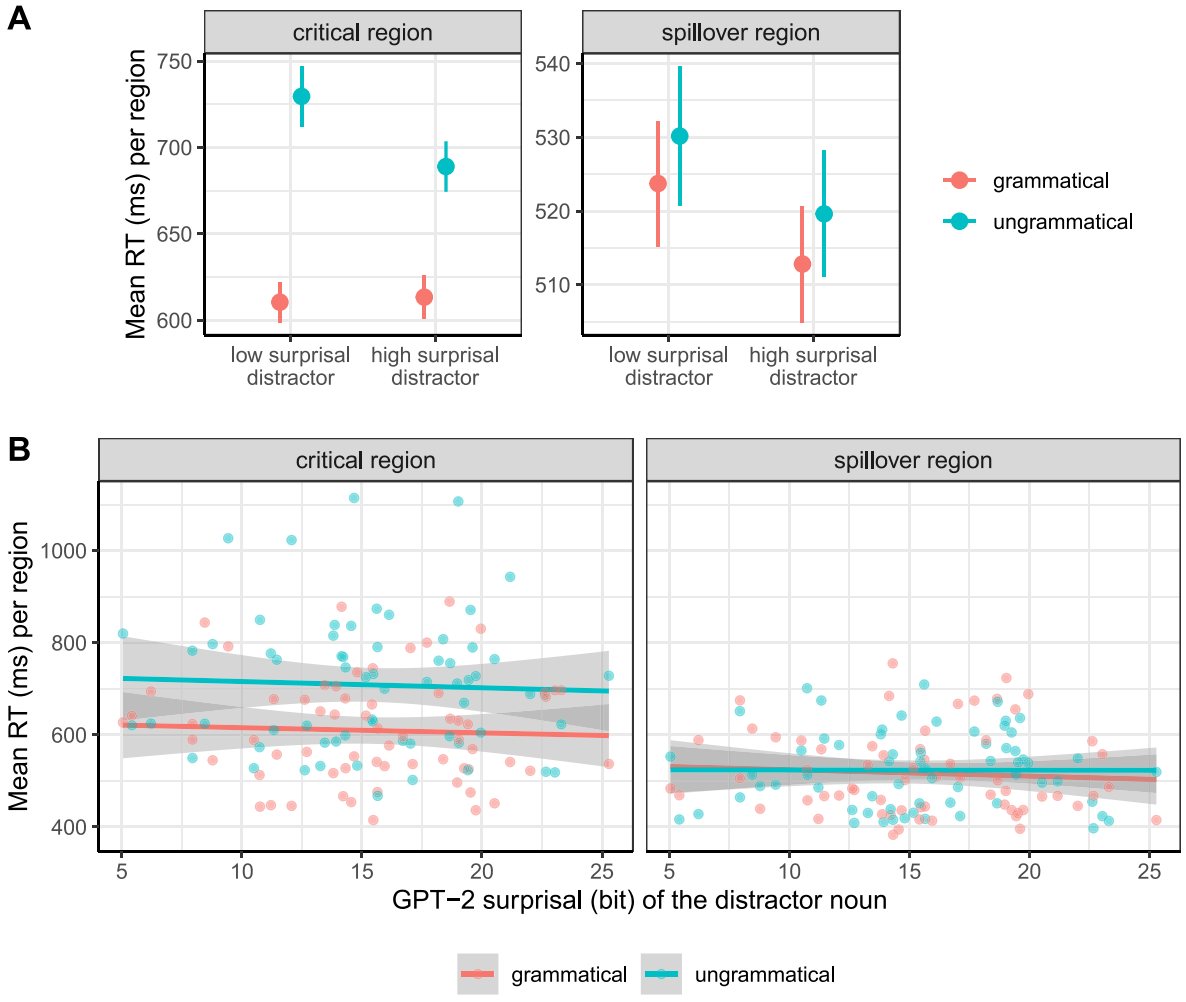

Fig. 7. Experiment 2 reading times in the critical and the spillover region. Target nouns are singular in all conditions. Distractors are plural in all conditions. *Panel A* corresponds to the binary surprisal (low vs. high) of the distractor noun; error bars represent standard errors of the mean. *Panel B* corresponds to the GPT-2 surprisal of the distractor noun, each point representing the aggregated reading time by item per condition.

Table 8
Experiment 2 statistical results on log-transformed reading times.

Binary surprisal	Critical region			Spillover region		
	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	−0.056	[−0.076, −0.037]	$P(\beta < 0) = 1$	0.003	[−0.015, 0.023]	$P(\beta > 0) = 0.638$
Surprisal	−0.012	[−0.026, 0.001]	$P(\beta < 0) = 0.965$	−0.009	[−0.022, 0.003]	$P(\beta < 0) = 0.933$
Gram × Surp	0.009	[−0.004, 0.022]	$P(\beta > 0) = 0.925$	−0.001	[−0.012, 0.010]	$P(\beta < 0) = 0.585$
GPT-2 surprisal	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	−0.057	[−0.078, −0.037]	$P(\beta < 0) = 1$	0.003	[−0.017, 0.022]	$P(\beta > 0) = 0.603$
Surprisal	−0.021	[−0.042, −0.001]	$P(\beta < 0) = 0.978$	−0.016	[−0.037, 0.004]	$P(\beta < 0) = 0.944$
Gram × Surp	0.007	[−0.010, 0.025]	$P(\beta > 0) = 0.805$	−0.003	[−0.016, 0.010]	$P(\beta < 0) = 0.705$

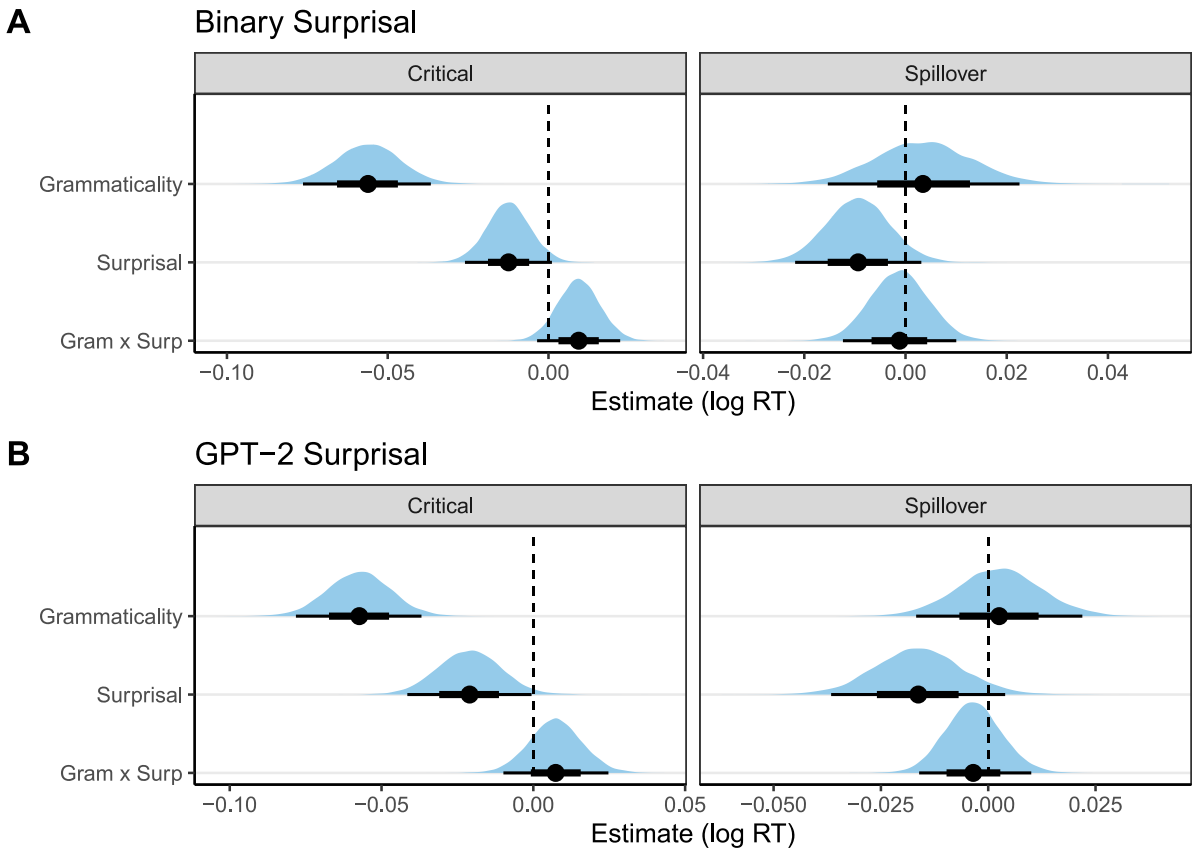

Fig. 8. Experiment 2 posterior distributions for log-transformed reading times. *Panel A*: Binary surprisal (low vs. high) of the distractor noun; *Panel B*: GPT-2 surprisal of the distractor noun. Black circles represent the mean of posterior estimates. Error bars represent 66% (thick) and 95% (thin) credible intervals.

Spillover region. The result in the spillover region does not qualitatively differ between the two surprisal measures. There is no longer evidence for a `Grammaticality` main effect in this region. There is still a main effect of distractor `Surprisal`, whereby more surprising distractors are associated with faster reading time at the retrieval site. No `Grammaticality` × `Surprisal` interaction is observed in the spillover region.

Discussion

In a simplified design, we only included plural distractors in Experiment 2. Since target nouns are in singular form in all conditions, this simplified design means that there is always some degree of interference from the distractor in terms of the number feature. The modulation on the attraction effect, therefore, is manifested as a modulation on the grammaticality illusion in this second experiment, such that a weaker grammaticality effect indicates stronger attraction from the plural distractor. As mentioned above, compared to Experiment 1, which looks into the effect of target noun surprisal on agreement attraction, we expect the surprisal of distractor noun to operate in the opposite direction, such that more surprising distractor noun should result in stronger attraction in Experiment 2. However, this is not what we observed. In Experiment 2, neither in acceptability judgments nor in online reading times did we observe a reliable effect of distractor surprisal on grammaticality, suggesting that distractor nouns of higher surprisal may not elicit stronger attraction to the retrieval of target nouns. At first glance, this lack of effect of distractor surprisal may seem at odds with our hypothesis of strategic memory allocation based on the assumption that both the target noun and the distractor noun share the same pool of working memory. That is, if more resources allocated to the encoding of the distractor noun result in fewer resources remained for the storage of the target noun, there should have been an effect of distractor surprisal as well, as opposed to the null result we actually observed in Experiment 2. This thus raises the question of how different linguistic units are structured in working memory to compete for the limited cognitive resources. We discuss our speculation below in General Discussion.

General discussion

In this paper, we proposed that linguistic units carrying more novel and unexpected information are prioritized for working memory resources, and are thus encoded with more robust memory representation against interference, a mechanism that we refer to as **strategic memory allocation**. We examined this hypothesis through the lens of the agreement attraction effect. In particular, we examined two predictions: (1) *target nouns* of higher surprisal should lead to *weaker* agreement attraction effect if they have more robust representation; (2) *distractor nouns* of higher surprisal should lead to *stronger* agreement attraction effect if they draw more memory resources from maintaining the representation of the target noun.

The first prediction was confirmed by our Experiment 1. As predicted, we found that more surprising target nouns are less disrupted by distractors, resulting in weaker agreement attraction effect. Moreover, this target noun surprisal effect on attraction magnitude is much more reliable when surprisal is measured as a fine-grained continuous variable from GPT-2 language model rather than as a binary categorization. This finding substantiates our hypothesis that novel and unexpected linguistic units are prioritized for working memory resources with more robust representation.

We also note two caveats in our result. First, the effect of target noun surprisal is only observed in online reading time data, but not in acceptability judgments, a divergence that needs to be reconciled in future work with a more detailed characterization of the linkage

between online and offline data. Second, although the reading time data shows the predicted pattern in the critical region, the effects in the spillover region are in an unexpectedly reversed direction. We have discussed our speculation about how this reversed effect may point to a later-stage reanalysis process for the originally misprocessed agreement features under attraction.

The second prediction, however, was not supported by our Experiment 2. Neither in the online reading time data nor in the offline acceptability judgments did we observe a statistically reliable effect of distractor surprisal on the magnitude of agreement attraction. As discussed below, this absence of informativity effect on distractors suggests that, rather than assuming a naive shared pool of resources for all the information, a more nuanced account of how information is structured and operated in both working and long-term memory is needed.

Absence of the effect of informativity on distractors

The absence of the distractor informativity effect in Experiment 2 raises the question about how different linguistic elements are structured in working memory. In order for the informativity of distractor nouns to influence the retrieval of the target noun, an important assumption is that these two linguistic units are structured in a way that they form a fair competition for the same pool of working memory resources.

Many previous studies, however, have observed that elements located in different linguistic contexts may not trigger memory interference to the same extent (Kim & Xiang, 2024; Van Dyke & McElree, 2011). For example, Van Dyke and McElree (2011) find that the configuration where the distractor is positioned between the target noun and the retrieval site (retroactive interference) induces stronger disruptive effect than the one where the distractor precedes the target noun (proactive interference). This observation implies that the linear order of memory elements does matters, possibly due to time-related decay (Barrouillet, Bernardin, & Camos, 2004; Lewis & Vasishth, 2005; Page & Norris, 1998; Portrat, Barrouillet, & Camos, 2008) and/or feature-overwriting exerted by subsequent information (Oberauer & Kliegl, 2006; Oberauer & Lange, 2008). This recency effect may have created a ceiling effect in our Experiment 2, such that the interference effect is already close to its maximum, leaving little room for further modulation by the surprisal of the distractor.

Another possibility is that informativity may have influenced the processing depth and the extent to which the underlying syntactic structure of the sentence is specified in memory. When the distractor is highly predictable and does not require much effort, the processing may be shallow, with many details of the syntactic structure underspecified (Ferreira, 2003a; Ferreira, Bailey, & Ferraro, 2002; Ferreira, Christianson, & Hollingworth, 2001; Ferreira & Patson, 2007; Sanford & Sturt, 2002; Tabor, Galantucci, & Richardson, 2004). On the contrary, an unexpected distractor noun may encourage the comprehender to perform deeper-level of processing, and to specify more detailed features of the syntactic structure. As a result, the less predictable distractor noun in our stimuli (as repeated below in (4)) may be better encoded with more robust representation as an embedded object inside the subject relative clause. This strengthened representation as an embedded object may create a stronger mismatch with the retrieval cue, which looks for a matrix subject. As a result, the attraction effect on the target noun may be canceled, even if the more surprising distractor noun is indeed prioritized and draws more working memory resources.

(4) *The monster [who chased the bald kids$_{embedded\ obj}$] seemingly were gone...

In fact, if the representation of the distractor is indeed strengthened as an embedded object, it provides another perspective to support our main hypothesis. Although our prediction for Experiment 2 focuses on how the informativity of the distractor may influence the representation of the target noun, it is also possible that the target noun's representation influences that of the distractor. Specifically, the structural feature of the target noun as a matrix subject interferes with the structural feature of the distractor noun, which is supposed to be an embedded object in the relative clause. According to our hypothesis, if the distractor noun is prioritized in working memory, it should receive a more robust representation as an embedded object, making it less susceptible to the interference from the target noun as a matrix subject. If this is the case, more unexpected distractor noun can potentially lead to weaker agreement attraction, too, due to the stronger mismatch with the retrieval cue, resulting in an empirical pattern that is in the same direction as our primary prediction in Experiment 1. In order to make more accurate empirical predictions, a model that quantifies the competition between the target noun and the distractor noun is needed in future work.

An interplay of memory, expectation, and learning

Both the relationship between expectation and memory, and the relationship between expectation and learning, have been long investigated in psycholinguistics. For the interplay between expectation and memory, on the one hand, much of the empirical work aims to tease apart the memory-based and the expectation-based mechanism in the processing of certain linguistic structures, for example, relative clauses (e.g., Grodner and Gibson 2005, Konieczny 2000, Levy 2013, Levy and Keller 2013, Nakatani and Gibson 2010, Ronai and Xiang 2023, Vasishth and Lewis 2006). On the other hand, some modeling studies attempt to account for both the surprisal effect and the locality effect within a unified sentence processing model (Demberg & Keller, 2009; Futrell, Gibson, & Levy, 2020; Rasmussen & Schuler, 2018). For example, grounded in the framework of surprisal theory (Levy, 2008a), the lossy-context surprisal model by Futrell, Gibson, and Levy (2020) integrates the mechanism of memory distortion into the prediction of upcoming linguistic units. Specifically, the prediction of the next word $p(w_i|r_{w_1 \ldots w_{i-1}})$ is based on the lossy memory representation r of the preceding context $w_1 \ldots w_{i-1}$ after memory distortion, instead of the true context as in the original surprisal theory $p(w_i \mid w_1 \ldots w_{i-1})$.

For the interplay between expectation and learning, the implicit learning account of priming and adaptation holds that learning goes hand-in-hand with processing over the life span of a language user, and that the adaptation of the mental model is driven by prediction error (Chang, Dell, & Bock, 2006; Jaeger & Snider, 2013; Xu & Futrell, 2024a). Specifically, the larger the prediction error, the stronger the learning effect, a pattern that is often empirically manifested as an inverse frequency effect in the priming literature (Bock, 1986; Ferreira, 2003b; Hartsuiker & Kolk, 1998; Kaschak, Kutta, & Jones, 2011; Scheepers, 2003). The prediction-driven learning has also been studied more generally in many statistical learning theories beyond the domain of language (Courville, Daw, & Touretzky, 2006; Elman, 1990; Rumelhart, Hinton, & Williams, 1986; Wagner & Rescorla, 1972).

Beyond how expectation interacts with memory and learning separately, the strategic memory allocation we proposed in this paper raises a potential mechanism that ties all three elements together. Upon encountering a linguistic unit with novel and unexpected information, the comprehender first puts much effort into encoding this unit. According to expectation-based theories, this additional processing effort arises due to the inconsistency between the bottom-up perceptual input and the top-down prediction generated by the mental model (Hale, 2001; Levy, 2008a). Presumably, what has been predicted depends on, first, what has been stored in memory, including both the long-term and the working memory (Futrell, Gibson, & Levy, 2020; Hahn, Degen, & Futrell, 2021; Ryskin & Nieuwland, 2023). Second, it also depends on how the mental model makes use of the information in memory to parse the perceptual input (Levy, 2008a; MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, 1996; Trueswell, Tanenhaus, & Garnsey,

1994). Therefore, when there is an inconsistency between bottom-up perceptual input and top-down predictions, it may serve as a signal to inform the comprehender that there is a need to update the mental model and the long-term knowledge, in order to make more accurate predictions in the future. As a result, more cognitive resources, such as attention and memory, are allocated to unpredictable linguistic units since they might be more important for updating the mental model.

At a more implementational level, the interplay of memory, expectation, and learning points to the predictive coding mechanism (Aitchison & Lengyel, 2017; Friston, 2005, 2010; Gagnepain, Henson, & Davis, 2012; Rao & Ballard, 1999). The theory holds that one of the main goals of the brain is to optimize its internal representation of what causes the sensory input by explaining away its prediction error, which is the difference between its prediction and the actual sensory input. In support of predictive coding, many studies focusing on perception find that it is the prediction error that is encoded in neural responses, and that the neural responses are enhanced when the sensory input is less expected (Blank & Davis, 2016; Murray, Kersten, Olshausen, Schrater, & Woods, 2002; Rao & Ballard, 1999; Sohoglu & Davis, 2016, 2020). This predictive coding in the brain is consistent with our proposal of strategic memory allocation, in the sense that unexpected sensory information that generates stronger prediction errors is prioritized in neural coding.

Memory efficiency and memory compression

The strategic memory allocation can be considered an efficient use of working memory resources, one of the major topics in working memory research both within and beyond the domain of language. An important mechanism for working memory efficiency is *memory compression*, or *abstraction*. The idea is that the encoded representation can be recoded into more compressed and abstract forms, dropping certain details and redundancies in the original sensory input (Bates & Jacobs, 2020; Brady et al., 2009, 2024; Brady & Tenenbaum, 2013; Christiansen & Chater, 2016). This corresponds to our intuition that sometimes what has been remembered in comprehension is not the full linguistic sequence in its exact form, but a gist of message with many details unspecified (Bradshaw & Anderson, 1982). The input data whose representation is more compressible can be more efficiently stored in memory, saving more memory space for other information (Bates, Lerch, Sims, & Jacobs, 2019; Brady et al., 2009). Compression can be implemented in a symbolic fashion such as chunking, with multiple low-level features being holistically represented by a high-level abstract one (Brady & Tenenbaum, 2013; Christiansen & Chater, 2016; Miller, 1956). More recently, compression is implemented with distributive representation. Under the framework similar to a variational autoencoder (VAE), Bates and Jacobs (2020) propose a model that encodes the input information through a latent memory representation of lower dimensionality that functions as memory bottleneck. A crucial objective of their model, then, is to find a representation, such that the original input can be decoded and reconstructed as accurately as possible, and that the performance of downstream cognitive tasks can be optimized.

Importantly, and highly relevant to the current study, memory compression is strongly sensitive to the statistical regularities in the input data. In vision working memory, as mentioned earlier in Introduction, the input stimuli is more compressible if the sensory features follow a highly correlated statistical distribution (e.g., Bates and Jacobs 2020, Brady et al. 2009, 2024, Brady and Tenenbaum 2013). In our result, we find that highly predictable target nouns elicit stronger agreement attraction effect. A potential mechanism from the perspective of memory compression is that linguistic units that are more statistically predictable from the context are more likely to be encoded in an abstract and compressed form, without putting much effort into encoding many of the morphosyntactic details in the original sensory input. This is because, according to our hypothesis, the original content of the more predictable units are easier to be reconstructed from their compressed forms. As a result, the underspecified morphosyntactic features on the predictable target noun may leave it more susceptible to the interference from distractors. This compression-based explanation for our findings in agreement attraction is consistent with many usage-based linguistics theories, where a sequence of linguistic units, if frequently used, becomes more likely to be automatically processed as a holistic construction, undergoing phonological reduction and morphosyntactic rigidification (e.g., Arnon and Snider 2010, Bybee 2006, Bybee, Perkins, and Pagliuca 1994, Fillmore, Kay, and O'Connor 1988, Goldberg 2003, Mansfield 2021, Tomasello 2005). It also echos the distinction between deep versus shallow processing mode mentioned above in Section "Absence of the effect of informativity on distractors" , in the sense that less predictable surprising information encourages the processor to go through deeper-level of processing and encode a representation that is less compressed.

The role of statistical regularities in memory compression also points to the interaction between working memory and long-term memory. From an information-theoretic perspective, an efficient way to use both long-term and working memory is probably to decorrelate the representations between them, such that things encoded in working memory bear little to low mutual information with those already stored in long-term memory. In other words, the coding strategy may be inefficient if working memory keeps re-encoding what has already been encoded in long-term memory. Therefore, a more efficient working memory encoding would be to prioritize information that is not yet encoded in long-term memory, which is in line with the strategic memory allocation that we have proposed in the current study. This idea of decorrelating working and long-term memory echos the previously proposed efficient coding strategy for neural populations (Atick, 1992; Barlow, 1961; Olshausen & Field, 1996; Rieke et al., 1995; Simoncelli & Olshausen, 2001; Wiechert et al., 2010), which claims that every possible combination of neural response levels should be equally likely to be used in order to maximize the computing resources of a neural population, resulting in a neural representation that is statistically decorrelated. It is also consistent with the predictive coding mechanism mentioned above, in the sense that the brain's strategy to encode prediction errors can be considered an efficient way to maximize the use of its neural resources by decorrelating the encoded information.

The interaction between long-term and working memory has also been discussed in the literature of chunking. Some studies propose that chunks are only stored in long-term memory, with the information in working memory stored in its uncompressed original form (Botvinick, 2005; Botvinick & Bylsma, 2005; Norris & Kalm, 2021; Norris, Kalm, & Hall, 2020; Norris et al., 2020). According to these studies, the memory performance for units that better match the statistical patterns in prior knowledge is enhanced since the chunks in long-term memory help to reconstruct the degraded representations in working memory, a process often referred to as *redintegration* in this line of research. Since the working memory representation itself is not compressed, these studies predict that redintegration only benefits units that have formed a chunk and the memory performance of other units should be unaffected. This prediction is indeed consistent with the absence of effect in our Experiment 2, where the distractor informativity does not affect the memory performance of the target noun. Besides redintegration, some other studies propose that statistically predictable units do form a compressed representation in working memory, and the compressed form serves as a content-free label that points to certain pieces of knowledge stored in long-term memory (Huang & Awh, 2018; Kanwisher, 1987; Thalmann et al., 2019). These studies predict that it takes additional effort to decode the compressed information by retrieving information from long-term memory through that content-free label. This prediction is actually consistent with a tendency in our result that low-surprisal target nouns are numerically more difficult to retrieve, an observation also reported in Hofmeister (2011).

In the end, contradictory as it might initially appear to our main proposal, it is worth noting that predictable information, when encoded

in a more compressed form to increase memory efficiency, often leads to improved memory performance in many empirical studies. That is, stimuli that are more consistent with prior knowledge (such as familiar or natural items), are easier to be encoded with greater accuracy in behavioral tasks (Bates & Jacobs, 2020; Blalock, 2015; Girshick, Landy, & Simoncelli, 2011; Jackson & Raymond, 2008; Xie & Zhang, 2017). In this sense, the memory representation of predictable information is robust, too. This is because, as mentioned earlier, predictable input is easier to be reconstructed based on prior knowledge. As long as the statistical structure of the experimental stimuli aligns with participants' long-term experience, as is indeed the case in many of these studies, behavioral performance should remain intact. In fact, this is why compressing predictable information improves memory efficiency: the reconstructed representation works well most of the time in long-term experience. However, in scenarios requiring more sensory details of the original input, which is possibly the case in the current study, unpredictable information should yield better behavioral performance since it is likely encoded in less compressed forms.

Conclusion

We have proposed that linguistic units with novel and unexpected information, despite their higher processing difficulty when first encountered, may actually be prioritized for working memory resources and receive a more robust memory representation against noise and interference in the communicative context, a mechanism that we refer to as **strategic memory allocation**. In support of this hypothesis, we have found that more surprising and informative target nouns indeed elicit stronger agreement attraction in online reading times. Moreover, the effect is much more reliable with the GPT-2 surprisal of target nouns, which is a more fine-grained predictability measure than a binary categorization. This finding substantiates strategic memory allocation by showing that the representation of surprising target nouns is indeed more robust and less susceptible to interference. Having established the effect on target nouns, we then manipulated the predictability of the distractor noun, assuming that more surprising distractors should draw more resources away from the target noun if they share the same pool of memory resources. However, no reliable modulation on agreement attraction was observed by the surprisal of distractor noun. This absence of distractor effect suggests the need for a more nuanced characterization of how information is structured and operated in memory. Our findings highlight an interplay of memory, predictive processing, and implicit learning, in the sense that more memory resources are allocated to unpredictable linguistic units since they might be more important for updating the comprehender's mental model. We also argued that the strategic memory allocation we proposed is an instantiation of efficient coding for both working and long-term memory. More broadly, by demonstrating that the limited working memory is efficiently and dynamically optimized for the relevant processing task, the current study provides a connection between sentence processing and the resource-rational analysis of human cognition in general.

CRediT authorship contribution statement

Weijie Xu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Richard Futrell:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Statistical models

Reading times

The prior of the Bayesian linear mixed-effect models on log-transformed reading times (RT) are specified as in (1):

(1) Prior on log RT
prior(normal(6, 1), class=Intercept)
prior(normal(0, 1), class = b)
prior(normal(0, 1), class = sigma)
prior(normal(0, 1), class = sd)
prior(lkj(2), class = cor)

Experiment 1 regression formulas on log-transformed RTs are in (2). The main analysis tests the `Grammaticality` × `Distractor` × `Surprisal` three-way interaction using the effect coding (that is, the sum contrast). The additional analysis tests the attraction effect within each grammaticality condition using the nested contrast, where the variable `Condition` has three levels contrasting: (a) plural and singular distractors within ungrammatical sentences (`Distractor` in ungrammatical); (b) plural and singular distractors within grammatical sentences (`Distractor` in grammatical); and (c) grammatical and ungrammatical conditions (`Grammaticality` main effect). Experiment 2 regression formulas on log-transformed RTs are in (3).

(2) Experiment 1 regression formula for log RT

- Effect Coding/Sum Contrast:
  logRT ∼ 1 + Grammaticality ∗ Distractor ∗ Surprisal + (1 + Grammaticality ∗ Distractor ∗ Surprisal | Subj) + (1 + Grammaticality ∗ Distractor ∗ Surprisal | Item)
- Nested Contrast:
  logRT ∼ 1+Condition ∗ Surprisal+(1+Condition ∗ Surprisal | Subj) + (1 + Condition ∗ Surprisal | Item)

(3) Experiment 2 regression formula for log RT

- logRT ∼ 1+Grammaticality ∗ Surprisal+(1+Grammaticality ∗ Surprisal | Subj) + (1 + Grammaticality ∗ Surprisal | Item)

The posterior probabilities were calculated using the function `as_draws_df()` in the `brms` package. The hypotheses tested are directional, since our prediction explicitly states that the attraction effect should be weaker when the target nouns are more surprising. We did not add the word length of the RT region of analysis into the statistical model since it does not qualitatively change the result.

Acceptability judgments

The acceptability judgment rates are re-scaled from $[0, 100]$ to $(0, 1)$ before statistical analysis, as introduced in the main article. The prior of the Bayesian linear mixed-effect models on the re-scaled acceptability judgments (AJ) in both experiments are specified as in (4).

(4) Prior on re-scaled AJ
prior(normal(0, 1), class=Intercept)
prior(normal(0, 1), class = b)
prior(normal(0, 1), class = sd)
prior(lkj(2), class = cor)

We used beta regressions that fit both the mean μ and the precision ϕ on the re-scaled AJ. The formulas for Experiment 1 and Experiment 2 are in (5) and (6) respectively:

(5) Experiment 1 regression formula for re-scaled AJ
bf(AJ ∼ 1 + Grammaticality ∗ Distractor ∗ Surprisal + (1 + Grammaticality ∗ Distractor ∗ Surprisal | Subj)+(1+Grammaticality ∗ Distractor ∗ Surprisal | Item), *phi* ∼ 1 + Grammaticality ∗ Distractor ∗ Surprisal+(1+Grammaticality ∗ Distractor ∗ Surprisal | Subj) + (1 + Grammaticality ∗ Distractor ∗ Surprisal | Item))

Table B.1
Experiment 1 full statistical result of the Bayesian beta regression on acceptability judgment rates.

	Binary surprisal			GPT-2 surprisal		
Mean μ	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.903	[0.796, 1.013]	$P(\beta > 0) = 1$	0.901	[0.791, 1.010]	$P(\beta > 0) = 1$
Distractor	0.061	[0.035, 0.086]	$P(\beta > 0) = 1$	0.060	[0.034, 0.087]	$P(\beta > 0) = 1$
Surprisal	−0.068	[−0.106, −0.029]	$P(\beta < 0) = 0.999$	−0.100	[−0.163, −0.031]	$P(\beta < 0) = 0.996$
Gram × Distr	−0.045	[−0.071, −0.019]	$P(\beta < 0) = 0.999$	−0.045	[−0.072, −0.018]	$P(\beta < 0) = 0.999$
Gram × Surp	−0.027	[−0.052, −0.003]	$P(\beta < 0) = 0.985$	−0.044	[−0.079, −0.009]	$P(\beta < 0) = 0.992$
Distr × Surp	−0.034	[−0.059, −0.009]	$P(\beta < 0) = 0.996$	−0.030	[−0.056, −0.004]	$P(\beta < 0) = 0.988$
Gram × Distr × Surp	0.004	[−0.021, 0.028]	$P(\beta > 0) = 0.615$	0.004	[−0.023, 0.031]	$P(\beta > 0) = 0.619$
Precision ϕ	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.041	[−0.051, 0.133]	$P(\beta > 0) = 0.813$	0.047	[−0.046, 0.142]	$P(\beta > 0) = 0.835$
Distractor	−0.058	[−0.095, −0.020]	$P(\beta < 0) = 0.998$	−0.055	[−0.095, −0.016]	$P(\beta < 0) = 0.997$
Surprisal	−0.040	[−0.079, −0.001]	$P(\beta < 0) = 0.979$	−0.055	[−0.110, −0.003]	$P(\beta < 0) = 0.981$
Gram × Distr	0.028	[−0.011, 0.067]	$P(\beta > 0) = 0.918$	0.025	[−0.015, 0.065]	$P(\beta > 0) = 0.887$
Gram × Surp	−0.045	[−0.082, −0.007]	$P(\beta < 0) = 0.99$	−0.059	[−0.112, −0.009]	$P(\beta < 0) = 0.99$
Distr × Surp	0.009	[−0.028, 0.046]	$P(\beta > 0) = 0.688$	0.026	[−0.015, 0.067]	$P(\beta > 0) = 0.896$
Gram × Distr × Surp	−0.045	[−0.087, −0.004]	$P(\beta < 0) = 0.984$	−0.041	[−0.086, 0.002]	$P(\beta < 0) = 0.97$

Table B.2
Experiment 2 full statistical result of the Bayesian beta regression on acceptability judgment rates.

Mean μ	Binary surprisal			GPT-2 surprisal		
	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.585	[0.448, 0.723]	$P(\beta > 0) = 1$	0.581	[0.438, 0.724]	$P(\beta > 0) = 1$
Surprisal	−0.118	[−0.162, −0.073]	$P(\beta < 0) = 1$	−0.172	[−0.237, −0.107]	$P(\beta < 0) = 1$
Gram × Surp	−0.004	[−0.041, 0.034]	$P(\beta < 0) = 0.573$	−0.026	[−0.077, 0.026]	$P(\beta < 0) = 0.839$
Precision ϕ	Estimate	95% CrI	Posterior probability	Estimate	95% CrI	Posterior probability
Grammaticality	0.027	[−0.073, 0.129]	$P(\beta > 0) = 0.705$	0.027	[−0.077, 0.131]	$P(\beta > 0) = 0.698$
Surprisal	−0.002	[−0.060, 0.056]	$P(\beta < 0) = 0.523$	0.009	[−0.058, 0.074]	$P(\beta > 0) = 0.606$
Gram × Surp	−0.039	[−0.101, 0.021]	$P(\beta < 0) = 0.898$	−0.030	[−0.106, 0.044]	$P(\beta < 0) = 0.783$

(6) Experiment 2 regression formula for re-scaled AJ
bf(AJ ~ 1 + Grammaticality * Surprisal + (1 + Grammaticality * Surprisal | Subj) + (1 + Grammaticality * Surprisal | Item), *phi* ~ 1 + Grammaticality * Surprisal + (1 + Grammaticality * Surprisal | Subj) + (1 + Grammaticality * Surprisal | Item))

Appendix B. Statistical results on acceptability judgments

See Tables B.1 and B.2.

Appendix C. Experiment 1 by-item and by-subject effects

See Fig. C.1.

Appendix D. Data exclusion

For participant-level exclusion based on the acceptability judgment rates (AJ) of infelicitous filler items, we tried to avoid an arbitrary threshold by removing participants whose mean AJ of infelicitous fillers is three standard deviations away from the mean across all the participants. For Experiment 1, this method was quite effective in removing participants who gave high ratings to infelicitous fillers, as shown. For Experiment 2, probably because of the smaller sample size, this method is less effective in removing participants who gave high ratings to infelicitous fillers. Therefore, for Experiment 2, we first removed participants whose mean AJ for infelicitous fillers is above 50/100, and then we further removed those whose mean AJ is beyond three standard deviations away from the mean across all the participants, as what we did for Experiment 1. After this step of exclusion, the AJ rate distribution for infelicitous fillers roughly aligns between both experiments.

In addition, we also did participant-level exclusion based on the mean reading times (RT) across the whole experiment. We performed this step of participant-level exclusion because we noticed on Prolific that there were some participants taking unreasonably long time to finish the experiment. This is especially the case for Experiment 2, as the mean RT per region is over 2 s for some participants. Again, we have checked that the distribution of mean RT across the whole experiment session roughly align between two experiments after this step of data exclusion.

For trial-level data exclusion, we checked that there is a relatively uniform distribution of removed RT data across all experimental conditions, indicating that our trial-level exclusion is reasonable and is not biased toward any specific experimental condition.

Table E.1
Interpretation of Bayes factors (Jeffreys, 1939).

Bayes factor (BF_{10})	Interpretation
> 100	Extreme evidence for M_1
30–100	Very strong evidence for M_1
10–30	Strong evidence for M_1
3–10	Moderate evidence for M_1
1–3	Anecdotal evidence for M_1
1	No evidence for M_1
1/3–1	Anecdotal evidence for M_0
1/10–1/3	Moderate evidence for M_0
1/30–1/10	Strong evidence for M_0
1/100–1/30	Very strong evidence for M_0
< 1/100	Extreme evidence for M_0

Appendix E. Bayes factors for critical region RT data

We calculated the Bayes factors (BF) for critical reading time effects in both experiments. For Experiment 1, we calculated the BF for the `Grammaticality` × `Distractor` two-way interaction and the `Grammaticality` × `Distractor` × `Surprisal` three-way interaction. We focus on the critical SPR region, where the effects were most reliably detected. For Experiment 2, we calculated the BF for the effect of `Grammaticality` × `Surprisal` interaction.

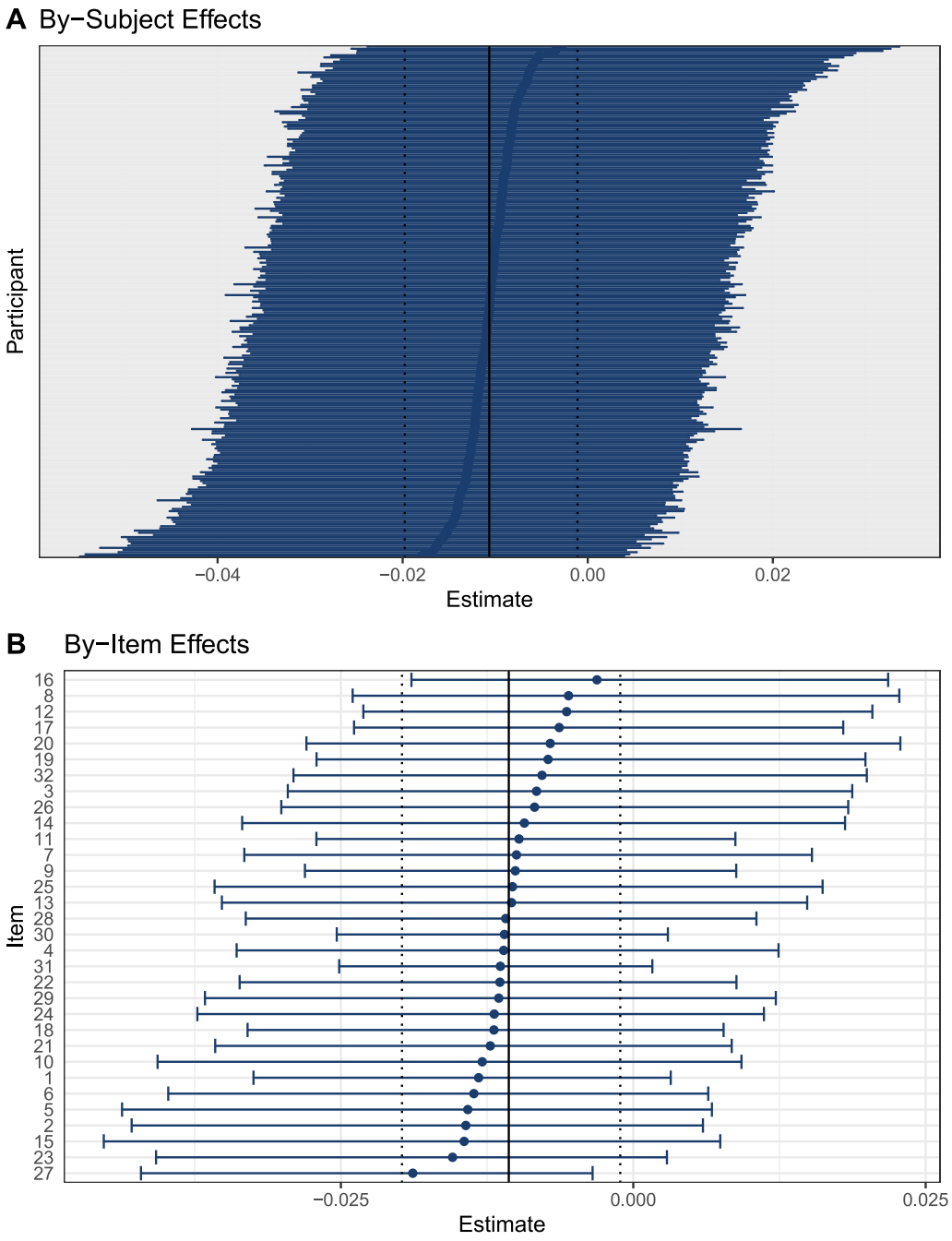

Fig. C.1. Experiment 1 by-subject and by-item effects of the Grammaticality × Distractor × GPT-2 Surprisal three-way interaction in the critical region. Solid vertical line shows the mean estimate of the fixed effect, with dashed vertical lines representing 95% credible intervals.

Bayes factors quantify the evidence for or against the effects of interest by comparing the alternative model (M_1) with the null model (M_0), analogous to the hypothesis testing using the likelihood ratio test in frequentist statistics (Gelman, Carlin, Stern, & Rubin, 2013; Schad, Nicenboim, Bürkner, Betancourt, & Vasishth, 2022). The analysis was implemented with `brms` package in R (Bürkner, 2017), and the marginal likelihoods were calculated through `bridge_sampler` (Gronau et al., 2017; Meng & Wong, 1996). Since the estimates of BFs can be unstable for models with complicated random effects, we simplified the random effects and only included by-item and by-participant random intercepts. We also increased the number of iterations to 20,000 per chain.

In Experiment 1, the regression formulas of null model M_0 and the full model M_1 are specified as in (7) and (8). When calculating BFs for the two-way interaction, we used the model with binary surprisal. In Experiment 2, the regression formulas of null model M_0 and the full model M_1 are specified as in (9).

(7) Experiment 1 model comparison for the `Grammaticality` × `Distractor` × `Surprisal` three-way interaction

- Null Model M_0:
  logRT ∼ 1 + Grammaticality ∗ Distractor + Grammaticality ∗ Surprisal + Distractor ∗ Surprisal + (1 | Subj) + (1 | Item)
- Full Model M_1:
  logRT ∼ 1 + Grammaticality ∗ Distractor ∗ Surprisal + (1 | Subj) + (1 | Item)

Table E.2
Experiment 1 Bayes factors (BF_{10}) in the critical region for the effects of `Grammaticality` × `Distractor` two-way interaction and `Grammaticality` × `Distractor` × `Surprisal` three-way interaction, with a range of standard deviations on the prior of the effects. For both interactions tested below, the null model includes all other interactions. BF_{10} quantifies the evidence for the alternative hypothesis (H_1) against the null hypothesis (H_0).$BF_{10} > 1$ favors H_1; $BF_{10} < 1$ favors H_0.

Prior	BF_{10} (Gram × Distr)	BF_{10} (Gram × Distr × Surp)	
		Binary surprisal	*GPT-2 surprisal*
$\mathcal{N}(\mu = 0, \sigma = 0.01)$	31.85	1.01	27.20
$\mathcal{N}(\mu = 0, \sigma = 0.05)$	9.59	0.24	10.31
$\mathcal{N}(\mu = 0, \sigma = 0.1)$	4.97	0.13	5.22
$\mathcal{N}(\mu = 0, \sigma = 0.5)$	0.99	0.02	1.05
$\mathcal{N}(\mu = 0, \sigma = 1)$	0.50	0.01	0.53

Table E.3
Experiment 2 Bayes factors (BF_{10}) in the critical region for the effect of `Grammaticality` × `Surprisal` interaction, with a range of standard deviations on the prior. $BF_{10} > 1$ favors H_1; $BF_{10} < 1$ favors H_0.

Prior	BF_{10} (Gram × Surp)	
	Binary surprisal	*GPT-2 surprisal*
$\mathcal{N}(\mu = 0, \sigma = 0.01)$	1.16	0.71
$\mathcal{N}(\mu = 0, \sigma = 0.05)$	0.35	0.18
$\mathcal{N}(\mu = 0, \sigma = 0.1)$	0.18	0.09
$\mathcal{N}(\mu = 0, \sigma = 0.5)$	0.04	0.02
$\mathcal{N}(\mu = 0, \sigma = 1)$	0.02	0.01

(8) Experiment 1 model comparison for the `Grammaticality` × `Distractor` two-way interaction

- Null Model M_0:
  logRT ~ 1 + Grammaticality * Surprisal + Distractor * Surprisal + Grammaticality : Distractor : Surprisal + (1 | Subj) + (1 | Item)
- Full Model M_1:
  logRT ~ 1 + Grammaticality * Distractor * Surprisal + (1 | Subj) + (1 | Item)

(9) Experiment 2 model comparison for the `Grammaticality` × `Surprisal` two-way interaction

- Null Model M_0:
  logRT ~ 1 + Grammaticality + Surprisal + (1 | Subj) + (1 | Item)
- Full Model M_1:
  logRT ~ 1 + Grammaticality * Surprisal + (1 | Subj) + (1 | Item)

As suggested by previous work, we obtained BFs with a range of standard deviations on the priors of the target effects as a sensitivity analysis (Schad et al., 2022). That is, the prior for log RT is set up as above in Appendix "Statistical Models", except that the standard deviation of the target effect in the full model M_1 is changed to the ones in Table E.2 and Table E.3 as a sensitivity analysis. Smaller prior standard deviations indicate that the models assume an effect close to zero, that is, a small effect size. Since Bayes factors compare the marginal likelihood of the observed data between M_0 and M_1, a smaller standard deviation for the prior of the target effect in M_1 makes the model more likely to capture the observed data if the effect size is indeed close to zero, and therefore makes the BF favors M_1 to a greater extent.

We followed the scale in Table E.1 proposed by Jeffreys (1939) to interpret Bayes factors. The BF result of Experiment 1 RT critical region is summarized in Table E.2. We obtained moderate to strong evidence for the `Grammaticality` × `Distractor` two-way interaction, and for the `Grammaticality` × `Distractor` × `Surprisal` three-way interaction with GPT-2 surprisal. This is true even for models assuming a 0.1 standard deviation in the prior, which should correspond to a decent effect size given that the model's dependent variable is RT in log scale. The BF result of Experiment 2 RT critical region is summarized in Table E.3. There is almost no evidence for a `Grammaticality` × `Surprisal` interaction with any surprisal measure.

Data availability

The trial-level data, analysis code, and the full list of stimuli are available at https://osf.io/e5dsv/.

References

Aitchison, L., & Lengyel, M. (2017). With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*(9), 578–582.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, *3*(2), 213–251.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans.. *Journal of Experimental Psychology: General*, *133*(1), 83.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory.. *Psychological Review*, *127*(5), 891.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*(2), 11–11.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.

Blalock, L. D. (2015). Stimulus familiarity improves consolidation of visual working memory representations. *Attention, Perception, & Psychophysics*, *77*, 1143–1158.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*(11), Article e1002577.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127.

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in english number agreement. *Language and Cognitive Processes*, *8*(1), 57–99.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, *2*(1).

Botvinick, M. (2005). Effects of domain-specific knowledge on memory for serial order. *Cognition*, *97*(2), 135–151.
Botvinick, M., & Bylsma, L. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(2), 351.
Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 165–174.
Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487.
Brady, T. F., Robinson, M. M., & Williams, J. R. (2024). Noisy and hierarchical visual memory across timescales. *Nature Reviews Psychology*, 1–17.
Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85.
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
Bruning, A. L., & Lewis-Peacock, J. A. (2020). Long-term memory guides resource allocation in working memory. *Scientific Reports*, *10*(1), 1–10.
Burchill, Z. J., & Jaeger, T. F. (2024). How reliable are standard reading time analyses? Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation. *Journal of Memory and Language*, *136*, Article 104494.
Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.
Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711–733.
Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. The University of Chicago Press.
Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic.. *Psychological Review*, *113*(2), 234.
Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, Article e62.
Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668.
Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.
Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review*, *119*(3), 480.
DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.
Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.
Demberg, V., & Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the Annual Meeting of the Cognitive Science Society*: *vol. 31*, (31).
Dempsey, J., Christianson, K., & Tanner, D. (2022). Misretrieval but not misrepresentation: A feature misbinding account of post-interpretive effects in number attraction. *Quarterly Journal of Experimental Psychology*, *75*(9), 1727–1745.
Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103.
Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.
Ferreira, F. (2003a). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.
Ferreira, V. S. (2003b). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, *48*(2), 379–398.
Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.
Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*, 3–20.
Ferreira, F., & Patson, N. D. (2007). The 'good enough'approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.
Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*(5), Article 056135.
Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 501–538.
Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *360*(1456), 815–836.
Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138.
Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), Article e12814.
Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.
Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.
Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.
Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.
Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity.. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: papers from the first mind articulation project symposium* (pp. 94–126). The MIT Press.
Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.
Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.
Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.
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* (pp. 10–18).
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.
Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.
Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., et al. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.
Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, *128*(4), 726.
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), Article e2122602119.
Hald, L. A., Steenbeek-Planting, E. G., & Hagoort, P. (2007). The interaction of discourse context and world knowledge in online sentence comprehension. Evidence from the N400. *Brain Research*, *1146*, 210–218.
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*.
Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, *110*, 70–104.
Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (pp. 75–86).
Hartsuiker, R. J., & Kolk, H. H. (1998). Syntactic persistence in dutch. *Language and Speech*, *41*(2), 143–184.
Hawkins, J. A. (1994). *A performance theory of order and constituency*. (73), Cambridge University Press.
Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.
Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Language and Cognitive Processes*, *26*(3), 376–405.
Hofmeister, P., & Vasishth, S. (2014). Distinctiveness and encoding effects in online sentence comprehension. *Frontiers in Psychology*, *5*, 1237.
Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & 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).
Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., et al. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, Article 104510.
Huang, L., & Awh, E. (2018). Chunking in working memory via content-free labels. *Scientific Reports*, *8*(1), 1–10.
Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31.
Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces.. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 556.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.
Jaeger, T. F., & Levy, R. (2006). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, *19*.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83.

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, Article 104063.

Jeffreys, H. (1939). *The theory of probability*. Oxford: Clarendon Press.

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*(2), 117–143.

Karimi, H., Diaz, M., & Ferreira, F. (2019). "A cruel king" is not the same as "a king who is cruel": Modifier position affects how words are encoded and retrieved from memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(11), 2010.

Karimi, H., Diaz, M., & Wittenberg, E. (2020). Sheer time spent expecting or maintaining a representation facilitates subsequent retrieval during sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society: vol. 2020*, (p. 2728). NIH Public Access.

Karimi, H., Diaz, M., & Wittenberg, E. (2023). Delayed onset facilitates subsequent retrieval of words during language comprehension. *Memory & Cognition*, 1–18.

Karimi, H., Swaab, T. Y., & Ferreira, F. (2018). Electrophysiological evidence for an independent effect of memory retrieval on referential processing. *Journal of Memory and Language*, *102*, 68–82.

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, *18*, 1133–1139.

Kim, S. J., & Xiang, M. (2024). Incremental discourse-update constrains number agreement attraction effect. *Cognitive Science*, *48*(9), Article e13497.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627–645.

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, Article 105159.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

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

Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234–243).

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence processing* (pp. 78–114). Psychology Press.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Levy, R., & Keller, F. (2013). Expectation and locality effects in german verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

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

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

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. *Cognition*, *233*, Article 105359.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, Article e1.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

Mansfield, J. (2021). The word as a unit of internal predictability. *Linguistics*, *59*(6), 1427–1472.

Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. (2023). Locally typical sampling. *Transactions of the Association for Computational Linguistics*, *11*, 102–121.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *99*(23), 15164–15169.

Nakatani, K., & Gibson, E. (2010). An on-line study of Japanese nesting complexity. *Cognitive Science*, *34*(1), 94–112.

Nicenboim, B., Schad, D., & Vasishth, S. (2024). An introduction to Bayesian data analysis for cognitive science. https://bruno.nicenboim.me/bayescogsci/. (Accessed 8 December 2024).

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

Norris, D., & Kalm, K. (2021). Chunking and data compression in verbal short-term memory. *Cognition*, *208*, Article 104534.

Norris, D., Kalm, K., & Hall, J. (2020). Chunking and redintegration in verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(5), 872.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626.

Oberauer, K., & Lange, E. B. (2008). Interference in verbal working memory: Distinguishing similarity-based confusion, feature overwriting, and feature migration. *Journal of Memory and Language*, *58*(3), 730–745.

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.

Oh, B.-D., Yue, S., & Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (volume 1: long papers)* (pp. 2644–2663). St. Julian's, Malta: Association for Computational Linguistics.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Paape, D., Avetisyan, S., Lago, S., & Vasishth, S. (2021). Modeling misretrieval and feature substitution in agreement attraction: A computational evaluation. *Cognitive Science*, *45*(8), Article e13019.

Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review*, *105*(4), 761.

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *Quarterly Journal of Experimental Psychology*, *69*(5), 950–971.

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

Portrat, S., Barrouillet, P., & Camos, V. (2008). Time-related decay or interference-based forgetting in working memory? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(6), 1561.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Rasmussen, N. E., & Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*, *42*, 1009–1042.

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *262*(1365), 259–265.

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition*, *209*, Article 104491.

Ronai, E., & Xiang, M. (2023). Memory versus expectation: Processing relative clauses in a flexible word order language. *Cognitive Science*, *47*(1), Article e13227.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Ryskin, R., & Nieuwland, M. S. (2023). Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*.

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, *158*, Article 107855.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386.

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2022). Workflow techniques for the robust use of Bayes factors. *Psychological Methods*.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205.

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), Article e2105646118.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), Article e2307876121.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 99–118.
Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.
Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54.
Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756.
Sohoglu, E., & Davis, M. H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. *eLife*, *9*, Article e58077.
Staub, A. (2024). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*.
Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.
Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215.
Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(1), 37.
Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
Troyer, M., Hofmeister, P., & Kutas, M. (2016). Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in Psychology*, *7*, 374.
Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*(4), 566–585.
Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*(3), 285–318.
Tung, T.-Y., & Brennan, J. R. (2023). Expectations modulate retrieval interference during ellipsis resolution. *Neuropsychologia*, Article 108680.
van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.
Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.
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.
Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 767–794.
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.
Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in pavlovian conditioning: Application of a theory. *Inhibition and Learning*, 301–336.
Wiechert, M. T., Judkewitz, B., Riecke, H., & Friedrich, R. W. (2010). Mechanisms of pattern decorrelation by recurrent neuronal circuits. *Nature Neuroscience*, *13*(8), 1003–1010.
Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (pp. 1707–1713).
Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.
Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672.
Xie, W., & Zhang, W. (2017). Familiarity increases the number of remembered pokémon in visual short-term memory. *Memory & Cognition*, *45*, 677–689.
Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721).
Xu, W., & Futrell, R. (2024a). A hierarchical Bayesian model for syntactic priming. In *Proceedings of the Annual Meeting of the Cognitive Science Society*: *vol. 46*.
Xu, W., & Futrell, R. (2024b). Syntactic dependency length shaped by strategic memory allocation. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP* (pp. 1–9).
Zehr, J., & Schwarz, F. (2018). PennController for internet based experiments (IBEX). http://dx.doi.org/10.17605/OSF.IO/MD832, Retrieved from.