Journal of Memory and Language 146 (2026) 104706

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Strategic resource allocation in memory encoding: An efficiency principle shaping language processing☆

Weijie Xu *, Richard Futrell

Department of Language Science, University of California, Irvine, United States of America

ARTICLE INFO

Dataset link: https://osf.io/yf4ca/

Keywords:
Strategic resource allocation
Working memory efficiency
Resource-rational
Dependency locality
Cross-linguistic
Sentence processing
Naturalistic data

ABSTRACT

How is the limited capacity of working memory efficiently used to support human linguistic behaviors? In this paper, we propose Strategic Resource Allocation (SRA) as an efficiency principle for memory encoding in sentence processing. The idea is that working memory resources are dynamically and strategically allocated to prioritize novel and unexpected information. From a resource-rational perspective, we argue that SRA is the principled solution to a computational problem posed by two functional assumptions about working memory, namely its limited capacity and its noisy representation. Specifically, working memory needs to minimize the retrieval error of past inputs under the constraint of limited memory resources, an optimization problem whose solution is to allocate more resources to encode more surprising inputs with higher precision. One of the critical consequences of SRA is that surprising inputs are encoded with enhanced representations, and therefore are less susceptible to memory decay and interference. Empirically, through naturalistic corpus data, we find converging evidence for SRA in the context of dependency locality from both production and comprehension, where non-local dependencies with less predictable antecedents are associated with reduced locality effect. However, our results also reveal considerable cross-linguistic variability, suggesting the need for a closer examination of how SRA, as a domain-general memory efficiency principle, interacts with language-specific phrase structures. SRA highlights the critical role of representational uncertainty in understanding memory encoding. It also provides a reinterpretation for the effects of surprisal and entropy on processing difficulty from the perspective of efficient memory encoding.

Introduction

Language processing in humans relies on working memory, a cognitive module known for its limited capacity to retain information (Baddeley, 1992; Fedorenko, Woodbury, & Gibson, 2013; Just & Carpenter, 1992). Under this limitation, a linguistic signal, once perceived, is at the risk of being lost, rapidly overwhelmed by the continual torrent of new inputs (Christiansen & Chater, 2016). Meanwhile, language use seems effortless, with sophisticated linguistic representations being dynamically encoded and decoded within milliseconds. This dual nature of working memory raises the question: how is the limited capacity of working memory *efficiently* used to support human linguistic behaviors?

In this paper, we propose *Strategic Resource Allocation* (SRA) as an efficiency principle for memory encoding in sentence processing. Specifically, working memory resources are dynamically and strategically allocated to prioritize novel and unexpected information. We argue that this efficiency principle, as a resource-rational theory (Gershman, Horvitz, & Tenenbaum, 2015; Lewis, Howes, & Singh, 2014; Lieder & Griffiths, 2020), naturally arises as the solution to a computational problem posed by two functional assumptions about working memory: its capacity is limited, and its representations are noisy. To examine this efficiency principle, we report three studies using naturalistic corpus data, where we demonstrate empirical support for strategic resource allocation through the lens of the locality effect in processing non-local syntactic dependencies.

Strategic Resource Allocation (SRA)

We first present the theoretical justification and existing empirical evidence for our proposal of Strategic Resource Allocation (SRA), drawing from the literature on sentence processing and psychophysics.

☆ This article is part of a Special issue entitled: 'Language Models & Psycholinguistics' published in Journal of Memory and Language.

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

https://doi.org/10.1016/j.jml.2025.104706
Received 27 October 2023; Received in revised form 28 August 2025; Accepted 17 October 2025

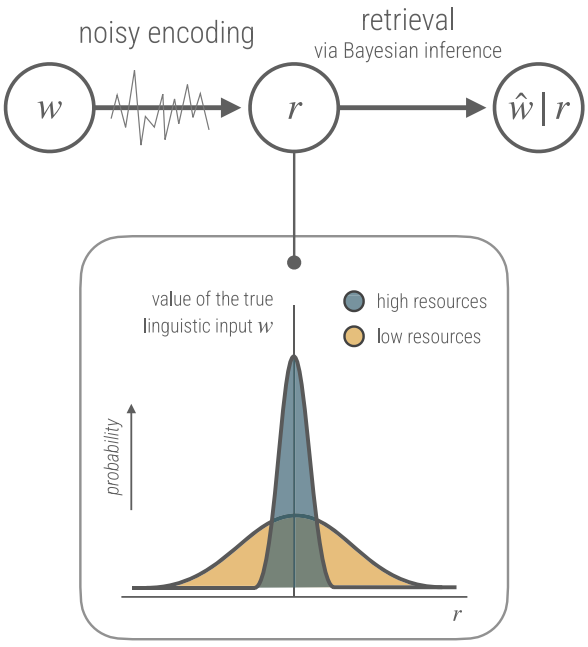


Fig. 1. Working memory processes under the probabilistic framework. The linguistic input w is encoded with noisy internal memory representation r. Higher memory resources results in sharper representation concentrated around the true input with less uncertainty. Memory retrieval can be considered an inference process reconstructing linguistic input based on noisy representation.

Theoretical proposal

We propose that working memory resources are strategically allocated in a way that prioritizes novel and unexpected information given the context, an efficiency principle that we refer to as *Strategic Resource Allocation* (SRA):

(1) Strategic Resource Allocation (SRA) in memory encoding:

Principle. Working memory resources are dynamically and strategically allocated in a way that prioritizes linguistic units that are unexpected and surprising given the context.

Core Prediction. The encoding of more surprising units is enhanced, resulting in more robust memory representations that are less susceptible to memory interference or decay.

This principle is in line with the resource-rational analysis of human mind (Gershman et al., 2015; Lewis et al., 2014; Lieder & Griffiths, 2020). Grounded in the bounded-rational approach to cognition (Anderson, 1990; Simon, 1955), resource-rational analysis aims to integrate the functional goals of a computational problem into the structural constraints of the underpinning cognitive architecture, providing a linkage between the computational-level and the algorithmic-level theories (Marr, 1982). In other words, instead of looking for an unbounded optimization, resource-rational analysis seeks to explain human behaviors under bounded rationality, that is, to identify an optimal solution that strikes the balance between maximizing the functional utility and adhering to the structural constraints of the cognitive system.

In this section, we will first outline the computational problem faced by working memory: to infer past information from uncertainty with maximal accuracy under the constraint of limited memory resources. We will then explain how SRA provides an optimal solution to this computational problem.

Inferring from uncertainty: A computational problem of working memory

As already mentioned, the resource-rational explanation for SRA is rooted in a computation problem posed by two functional assumptions about working memory. First, the capacity of working memory is limited. Although the exact nature of this limitation is still under debate, recent models in some non-linguistic domains have shifted from a discrete slot representation (Cowan, 2001; Luck & Vogel, 1997; Miller, 1956; Pashler, 1988) towards a continuous resource-based representation, where the limited resources can be flexibly allocated across the encoded information (Bates & Jacobs, 2020; Brady, Störmer, & Alvarez, 2016; Brady & Tenenbaum, 2013; Jakob & Gershman, 2023; Ma, Husain, & Bays, 2014; Sims, 2016; Sims, Jacobs, & Knill, 2012; van den Berg & Ma, 2018; van den Berg, Shin, Chou, George, & Ma, 2012).

Second, memory representation is full of noise and uncertainty, with unpredictable corruption in the veridical forms of sensory input, resulting in distorted representations that undermine behavioral performance such as inaccurate recall and illusive comprehension (Brady, Robinson, & Williams, 2024; Ferreira, Bailey, & Ferraro, 2002; Gibson, Bergen, & Piantadosi, 2013; Levy, 2008b; Ma et al., 2014). This uncertainty is often represented under the probabilistic framework. As shown in Fig. 1, when a linguistic input w is received, it can be encoded into an internal representation through certain memory model $r = M(w)$. This r is a probabilistic distribution centered around the true value of that input, such that the true input bears the highest probability in the encoding distribution compared to other alternatives.[1] Given this noisy representation, the true state of a past input is inaccessible, and memory recall, decoding, or retrieval, is effectively an inferential process that reconstructs past input from uncertainty using the statistical structure of long-term knowledge. The results of this process are often mathematically characterized using Bayes' rule (e.g., Bays, Schneegans, Ma, & Brady, 2024; Futrell, Gibson, & Levy, 2020; Gibson et al., 2013; Levy, 2008b; Ryskin et al., 2021):

$$p(\hat{w} \mid r) \propto p_M(r \mid w)p(w). \tag{1}$$

The equation describes a rational Bayesian decoder which infers the input from a specific memory representation r integrating prior knowledge $p(w)$, yielding a posterior distribution $p(\hat{w} \mid r)$. Then, marginalizing over all possible values of r, the distribution on the reconstructed word given the true input w^* is

$$p(\hat{w} \mid w^*) = \int p(\hat{w} \mid r)p(r \mid w^*)dr. \tag{2}$$

In this inferential process, inputs that are more probable in the prior are more likely to be accurately reconstructed, resulting in higher retrieval accuracy. See "Appendix A.1" for detailed mathematical formalization of this probabilistic memory encoding and retrieval process.

The two assumptions above naturally give rise to the following optimization challenge: *how to maximize memory accuracy under the constraint of limited resources?* At the core of this challenge lies an *efficiency* problem for two reasons. First, there is a functional goal, which is memory accuracy, against which the working memory performance is evaluated. Such a functional nature situates the current proposal under the rationalist approach to human mind. Second, working memory has internal constraints, in the sense that there is something it cannot achieve due to the cost from its own structure. Without such constraints, there would be no reason to look for an *efficient* implementation of a functional goal. The acknowledgment of system-internal

[1] In many psychophysics studies, the encoded representation is assumed to be a specific stimulus value that is generated from certain probabilistic distribution. Here in our work, we take a different assumption and postulate that what has been encoded, instead of a specific stimulus value, is the distribution itself, either through sampling (Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023) or through probabilistic population codes (Ma, Beck, Latham, & Pouget, 2006).

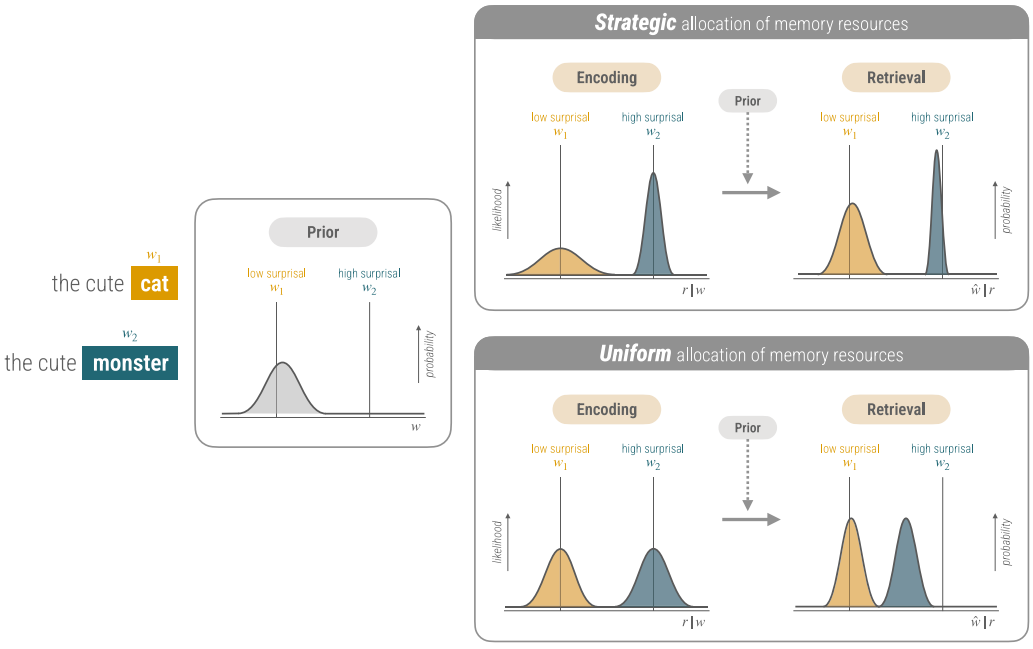


Fig. 2. Conceptual framework of strategic resource allocation (SRA) versus uniform resource allocation in memory encoding. Compared to uniform allocation, SRA holds that more surprising units receive more memory resources, therefore less representational uncertainty in its encoding distribution. Upon retrieval, more surprising units are less reconstructable based on prior compared to less surprising ones. But SRA leads to lower retrieval error overall by improving the reconstruction of high-surprisal units with minimal loss to the retrieval accuracy on low-surprisal units.

cost, therefore, further situates the current proposal under resource-rationality.[2] Next, we will explain how SRA provides a principled solution to the computational challenge outlined above.

SRA as a resource-rational solution

First of all, an important assumption we make is that the precision of the encoded distribution is proportional to the amount of memory resources allocated to an input unit. That is, as illustrated in Fig. 1, more resources allocated to encode w results in sharper distribution with less uncertainty, such that more probability mass is concentrated around the true input value w. Despite its lack of attention in the field of sentence processing, this assumption has been widely entertained in the literature of psychophysics (e.g., Bates & Jacobs, 2020; Bays, Catalao, & Husain, 2009; Bays et al., 2024; Ma et al., 2014).[3] As shown below, such a relationship between allocated resources and representational uncertainty lays the foundation for the derivation of SRA.

Before demonstrating the rationale behind SRA, let us first consider a naive strategy in which memory resources are uniformly distributed across all linguistic units regardless of the statistical structure in the context (Fig. 2, bottom panel). Under this uniform distribution, each input unit will receive an encoding distribution with identical precision. However, under the influence of prior, inputs that are more surprising under the prior will be less reconstructable, and the retrieval distribution will be more drawn towards the prior. Due to this difference in reconstructability, the same encoding distribution would yield different retrieval accuracy, disproportionally exerting impact on high surprisal inputs, reducing their retrieval accuracy more significantly than low surprisal ones. Therefore, a uniform distribution of memory resources is not the most efficient way to go for memory encoding, leaving substantial room for the improvement of overall retrieval accuracy.

Now, consider SRA, the strategic allocation of memory resources (Fig. 2, top panel). Recall that the idea is to strategically allocate more resources on linguistic units of higher surprisal *a priori*. That means, the prioritized more surprising units will receive sharper encoding with higher precision. This asymmetric allocation of resources is more efficient than the uniform strategy described in the last paragraph, since it achieves higher accuracy on average across inputs. By sacrificing a slight reduction in retrieval accuracy for low surprisal units, significant gains can be achieved for high surprisal ones by preventing these irreconstructable units from being distorted in the first place. Put simply, when only a limited number of linguistic units can be encoded with minimal distortion, it is more important to encode the more surprising and less reconstructable ones (See "Appendix A.2" for mathematical derivation).[4]

Beyond memory, SRA aligns with theories such as predictive coding, free energy principle, and implicit learning. At the neural level, the predictive coding mechanism (Aitchison & Lengyel, 2017; Blank &

[2] In this paper, we simply represent the internal cost of working memory as a computational bound (i.e., the total amount of available memory resources) within which a given task such as memory retrieval is imperfectly optimized, an approach that has been termed *bounded optimality* (Icard, 2023).

[3] One of the biggest challenges to apply such an encoding distribution in the domain of language is the specification of hypothesis space. In psychophysics, the hypothesis space is usually a quantitative spectrum that can be objectively specified based on certain physical features. But for language, the linguistic inputs are discrete units. Nowadays, with the advance of modern NLP techniques, this challenge has been significantly mitigated given the distributive word representations such as word embeddings. However, it is still nontrivial work to figure out what kind of probabilistic distribution should be applied to word embedding space. See "Appendix A.1" for a Gaussian approximation to the probabilistic memory processes.

[4] By having an encoding strategy that optimizes faithful reconstruction of the true input, an implicit assumption we made is that the input signal is considered error-free. If the input itself contains errors (e.g., when there are speech errors produced by the speaker), a reconstruction that is strongly influenced by the prior may actually be preferred so that the signal errors can be corrected. How to deal with the errors in input signals is an important online processing task. But in the current proposal, we choose to analyze working memory as a system of information storage, whose main goal is to accurately encode and decode the information it receives (cf. Hasson, Chen, & Honey, 2015).

Davis, 2016; Gagnepain, Henson, & Davis, 2012; Murray, Kersten, Olshausen, Schrater, & Woods, 2002; Rao & Ballard, 1999; Sohoglu & Davis, 2016, 2020) and the free energy principle (Friston, 2005, 2010; Gershman, 2019) hold that the brain seeks to minimize its prediction error, or surprise, as a way to optimize its internal model of the external environment. This principle is implemented by encoding prediction errors rather than the raw sensory input in neural signals. At the behavioral level, implicit learning theories often hold that learning is error-driven, with considerable empirical support showing that larger prediction errors lead to greater learning effect (Bock, 1986; Chang, Dell, & Bock, 2006; Courville, Daw, & Touretzky, 2006; Elman, 1990; Ferreira, 2003; Hartsuiker & Kolk, 1998; Jaeger & Snider, 2013; Rumelhart, Hinton, & Williams, 1986; Scheepers, 2003; Wagner & Rescorla, 1972; Xu & Futrell, 2024). Taken together, all these theories delivered a similar implication for our proposal in the domain of working memory. That is, when predictions conflict with the actual perceptual input (that is, when there is high surprisal input), it signals the need for comprehenders to update their mental model in order to make more accurate predictions in the future. Given this critical role of more surprising linguistic units in refining the mental model, it is reasonable to allocate more memory resources to them.

Predictability–precision trade-off

SRA can also be construed as a *predictability–precision trade-off*, as illustrated in Fig. 3. Imagine there are multiple linguistic units to be stored in working memory, with the goal of minimizing their total retrieval error (or, maximizing the retrieval accuracy). In most cases, for each unit, higher memory resources suggest higher encoding precision, which lead to lower retrieval error.[5] However, given a fixed amount of resources, allocating more to one unit necessarily reduces what is available to others. Therefore, the gain of retrieval accuracy for one unit necessarily comes with the loss for others. Consequently, to minimize total retrieval error, the *optimal* distribution of a fixed amount of resources should balance the gain and the loss in retrieval accuracy across units.

Where, then, does the balance hold? The intuition is illustrated in Fig. 3. First, let us start from the uniform distribution of resources, where both high- and low-surprisal inputs receive the same amount of resources, and are thus encoded with the same degree of precisions (i.e., gray dots in Fig. 3). Importantly, the retrieval error decreases faster for high-surprisal input. Therefore, at this point of uniform distribution of resources, there is a momentum to redistribute more resources to high-surprisal input. This is because such a redistribution towards high-surprisal input reduces the error faster. In other words, given a fixed amount of resources, the pressure to lower the overall retrieval error will push the encoding precision of high-surprisal input to increase from a uniform distribution of resources, and vice versa for low-surprisal one. This redistribution continues until there is balanced marginal effects, that is, when the slopes of the two error functions are equal, as in Fig. 3 (see "Appendix A.2" for details).

Precision, accuracy, and robustness

It is important to point out that the key prediction of SRA is not simply about the mean accuracy of memory retrieval. As shown in Fig. 2 (top panel), both predictable and unpredictable units can achieve relatively high accuracy with respect to the mean of retrieval distribution. For predictable input, this is supported by prior knowledge; for unpredictable input, this is achieved by more precise encoding representation. In fact, one of the critical consequences of SRA is that the retrieval mean accuracy should remain approximately similar across different linguistic units.[6]

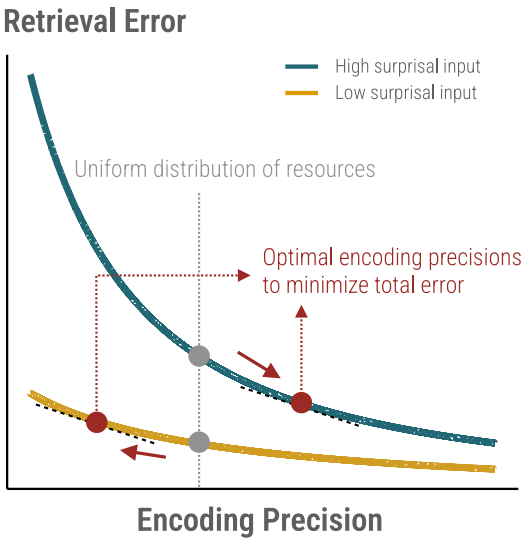


Fig. 3. Retrieval error as a function of encoding precision for high-surprisal and low-surprisal inputs. The optimal encoding strategy balances the potential gain and loss in retrieval accuracy across linguistic inputs.

But what *does* differ is the *precision*, or *uncertainty* in the representation. By allocating more resources, more surprising linguistic units are encoded with lower uncertainty (see the sharper distributions with high resources in Fig. 2). We thus propose that the uncertainty in the memory representation, rather than being linked to retrieval accuracy, is more directly related to *memory robustness*. That is, memory representation of higher robustness is less susceptible to the interference from other elements in memory. It of course remains a debatable question what the linking hypothesis is for representational uncertainty, but the linkage between uncertainty and robustness gives us a working hypothesis that is readily testable, as shown below in the rest of this paper. We will return to this point later in General Discussion "The role of representational uncertainty".

Our theoretical framework of SRA alludes to three effects on memory encoding precision:

(2) Three predicted effects of strategic resource allocation on encoding precision

   a. *Effect of input surprisal*
      Surprising linguistic units bear higher encoding precision, resulting in more robust memory representation against interference.
   b. *Effect of memory constraint*
      More available memory resources result in higher encoding precision overall.
   c. *Effect of prior precision*
      Precision of prior prediction does not necessarily increase or decrease encoding precision.

[5] In some borderline cases, where the input word is too close to the prior prediction and the prior precision is too unreliable, the retrieval error may not monotonically decrease with increasing encoding precision. However, as shown in "Appendix A.2", in either case, the strategic resource allocation should still hold in the sense that more resources should be allocated to high surprisal units in order to minimize the expected total error.

[6] SRA does not necessarily predict the retrieval mean accuracy to be an absolute constant across all linguistic units. In fact, the accuracy for low surprisal units may still be higher than high surprisal ones in Bayesian inference. However, due to SRA, this difference can be reduced compared to a naive encoding strategy such as uniform resource allocation.

The most important prediction of SRA is (2a), the effect of input surprisal. As outlined earlier, the optimal strategy to minimize overall retrieval error is to allocate more memory resources to encode surprising input. This strategy results in higher precision in the representation, thus higher memory robustness against interference. This is the critical prediction that we are going to examine empirically in the current study.

For the effect of memory constraint (2b), SRA predicts that more available memory resources results in higher encoding precision in general. This will also lead to higher memory robustness and more accurate retrieval overall.

For the precision of prior prediction (2c), its effect on encoding precision is in fact less straightforward. When the true input is very close to the prior prediction, it is indeed possible that less uncertainty in the prior can better support memory retrieval, thus less precise encoding is needed. However, when the true input is far from the prior prediction, it is not necessarily the case that more precise prior can still support better retrieval. We will discuss this effect in more detail and its implication for the effect of prediction entropy on processing difficulty in General Discussion "Processing difficulty as encoding difficulty: Reinterpreting the effect of surprisal and entropy".

Some existing empirical evidence

A dynamic similar to strategic resource allocation (SRA) is observed in the resource-rational model of sentence processing by Hahn, Futrell, Levy, and Gibson (2022). Grounded in the framework of lossy-context surprisal (Futrell, Gibson, & Levy, 2020), their model involves a contextual representation that represents only those words that are most useful for a downstream next-word prediction task. Their model predicts that function words, which are mostly predictable from the linguistic context, are more likely to undergo decay. In fact, our proposal of strategic resource allocation and lossy-context surprisal theory form two sides of the same coin in many aspects. We will discuss the relationship between these two theories in General Discussion Section "Relationship with lossy-context surprisal".

Studies focusing on memory retrieval mechanisms find that linguistic units of higher semantic complexity can be more easily retrieved in later stages of processing despite the initial encoding difficulty, implicating an enhanced accessibility for informative content from the model-theoretic perspective of informativity (Hofmeister, 2011; Hofmeister & Vasishth, 2014; Karimi & Ferreira, 2016; Troyer, Hofmeister, & Kutas, 2016). However, the exact cognitive underpinning for this empirical observation is still debatable (Hofmeister & Vasishth, 2014; Karimi, Diaz, & Wittenberg, 2023), and there is lack of clear empirical evidence for whether this effect can be extended to the information-theoretic view of informativity based on probabilistic prediction (Shannon, 1948). In spite of these unsettled issues, as a preliminary evidence from the existing literature, the effect of facilitated retrieval for semantically complex units aligns with our rational account outlined above, in the sense that the enhanced accessibility associated with informative units results from the prioritized resources allocated to their encoding.

Recently, SRA is more directly examined by Xu and Futrell (2025) through the lens of the agreement attraction effect in English. As shown below, even though the sentences in (3) are ungrammatical in English due to the mismatch of number feature between the subject head noun and the main verb, they are often perceived grammatical by native speakers due to the interference from the distractor noun in between, which shares the number feature with the ungrammatical main verb. In Xu and Futrell (2025), by manipulating the surprisal of the subject head noun through a prenominal adjective, they find that, compared to more surprising subject head nouns (e.g., cute *monster*), less surprising ones (e.g., evil *monster*) lead to stronger agreement attraction effect, such that the processing of the main verb is less susceptible to the interference from the distractor noun. They interpret the result as evidence for an enhanced memory representation of more surprising linguistic units against memory interference.

(3) a. *The evil *monster* who chased the kids seemingly *were* gone before the sunset. [low surprisal]
b. *The cute *monster* who chased the kids seemingly *were* gone before the sunset. [high surprisal]

In visual working memory, statistical regularities in long-term knowledge have been shown to shape memory performance. Despite the fact that items more consistent with prior knowledge are easier to be encoded with lower neural activity and enhanced behavioral performance (Bates & Jacobs, 2020; Blalock, 2015; Girshick, Landy, & Simoncelli, 2011; Jackson & Raymond, 2008; Xie & Zhang, 2017), some recent studies indeed observe that, in later stages of processing, these familiar items are de-prioritized to save more resources for the processing of novel ones (Brady et al., 2024; Bruning & Lewis-Peacock, 2020; Hedayati, O'Donnell, & Wyble, 2022; Kowialiewski, Lemaire, & Portrat, 2022). For example, in a delayed-estimation task, Bruning and Lewis-Peacock (2020) ask participants to first memorize and then recall the exact locations of six colored balls on a circle after a brief delay. Before the task, a sub-area on the circle has been previously illustrated to certainly contain the ball with a specific color (e.g., red ball) as a prior information. Their critical finding is that colors not included in the prior information (e.g., non-red balls) have lower recall accuracy when positioned closer to that sub-area, suggesting that memory resources have been shifted away from the prior area to prioritize other areas where novel information is more likely to appear.

SRA and dependency locality

The empirical focus of this paper to examine SRA is the *locality effect* in sentence processing, which has been considered a representative example of the efficient use of working memory resources. In this section, we will first introduce the empirical background of dependency locality effect. Then, we will present the empirical predictions of SRA in the context of dependency locality.

Dependency locality

Consider the sentence pair in (4). In (4a), codependents in the subject–verb dependency are adjacent to each other, whereas in (4b), there is additional linguistic material in between:

(4) a. The *monster approached* the princess...
b. The *monster* who stayed in the tower *approached* the princess...

The Dependency Locality Theory (DLT) (Gibson, 1998, 2000) holds that the formation of the non-local structures is constrained by the limited capacity of working memory. Specifically, as dependency distance increases, there is a higher memory cost to store the incomplete dependency as well as a higher integration cost to compute the new structural representation when the other codependent is encountered. In support of DLT, increased processing difficulty is often associated with structures that have longer dependency distance (e.g., Bartek, Lewis, Vasishth, & Smith, 2011; Ford, 1983; Gordon, Hendrick, & Johnson, 2001; Grodner & Gibson, 2005; King & Just, 1991; Miller & Isard, 1964; Traxler, Morris, & Seely, 2002; Yngve, 1960).[7] Similarly, in the resolution of structural ambiguity where a constituent has multiple potential attachment sites, there is a tendency for comprehenders to prefer the structure with local attachment (Frazier & Fodor, 1978; Gibson, Pearlmutter, Canseco-Gonzalez, & Hickok, 1996; Pearlmutter & Gibson, 2001).

[7] There is actually an anti-locality effect often found in some head-final dependencies, which is considered to be better explained by an expectation-based mechanism (Konieczny, 2000; Levy & Keller, 2013; Nakatani & Gibson, 2010; Vasishth & Lewis, 2006).

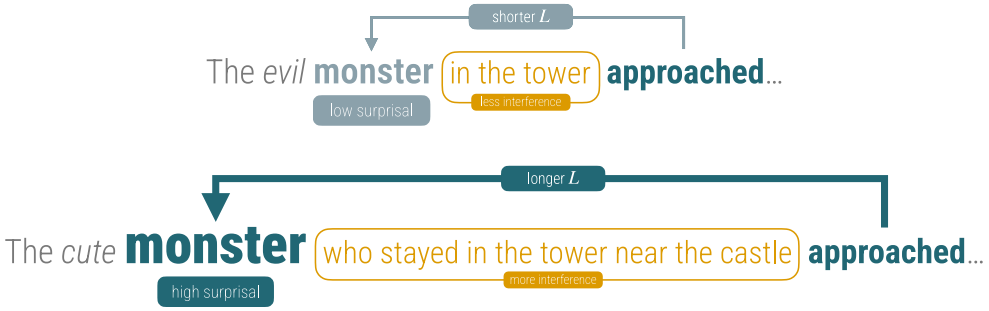


Fig. 4. Empirical prediction of strategic resource allocation in dependency locality. High-surprisal antecedents are more tolerable to longer dependency length.

Due to the memory constraint involved in processing non-local dependencies, an efficiency principle for language structure should be that linguistic units connected in a syntactic dependency tend to stay close in linear order. This locality principle is evidenced by cross-linguistic word-order patterns (Ferrer-i Cancho, 2004; Futrell, Levy, & Gibson, 2020; Futrell, Mahowald, & Gibson, 2015; Hawkins, 1990, 1994, 2004; Liu, 2008, 2021; Liu & Wulff, 2023; Temperley & Gildea, 2018) (cf. Liu, 2020), and has been argued to explain typological patterns such as the consistency in head direction, the contiguity of constituents, and the asymmetry of short-before-long versus long-before-short between head-initial and head-final languages (Futrell, Levy, & Gibson, 2020; Hawkins, 1994, 2004).

More recently, some studies propose a generalization from dependency locality to information locality, where any pair of linguistic units with high co-occurrence statistics, no matter whether they are in the same syntactic dependency or not, should stay close in linear order (Futrell, 2019; Futrell, Gibson, & Levy, 2020; Hahn, Degen, & Futrell, 2021; Hahn & Xu, 2022). Compared to previous work, these studies highlight the role of predictive processing, pointing out an interaction between the memory-based and the expectation-based mechanisms. Specifically, under the framework of Surprisal Theory (Hale, 2001; Levy, 2008a), the processing difficulty of a linguistic unit is proportional to how well it is predictable from the memory representation of the past input, which is prone to memory loss and distortion. The locality effect, as an efficient use of working memory, suggests that linguistic units carrying the most relevant information to predict the current one should stay in the recent past before they are forgotten.

These locality principles depend on the precise nature of working memory. Therefore, beyond the general capacity-based constraint proposed by DLT, it remains an open question how far this efficiency account can go with more and more realistic and detailed characterization of the nature of working memory constraints. Moreover, the existing discussion in the literature rarely addresses efficiency in processing per se. In other words, it is possible that memory limitations make language users not only actively choose a sentence form that is easier to process, but also develop an efficient processing strategy to better handle the information they passively receive.

The current study

We examine SRA in the context of dependency locality through naturalistic corpus data. If working memory resources are indeed dynamically and strategically allocated such that novel and unexpected information is prioritized, we predict that antecedents (i.e., left codependents) that are more surprising should receive sharper encoding with less uncertainty. The consequence of this is that memory for more surprising antecedents is enhanced, making their representations less susceptible to memory decay and interference before they need to be re-accessed at the other side of the dependency. Therefore, as illustrated in Fig. 4, more surprising antecedents should be able to tolerate longer dependency length, resulting in a reduced locality effect. We approach this prediction from both production (Study 1) and comprehension (Study 2a and 2b).

There are two terminological clarifications. First, we adopt a relatively broad interpretation of the term *memory encoding* in this article, focusing on the representational aspect of working memory mechanisms. Second, the term *resources* refers to any quantity that is limited and costly to use for better cognitive performance. Given the ongoing debate about the exact nature of working memory resources (Bays et al., 2024; Ma et al., 2014), we choose to restrict the use of this term to its abstract sense.

To preview our results, we find converging evidence from both production and comprehension that unexpected information is encoded with enhanced robustness against decay and interference. In Study 1, which focuses on production data, we observe that more surprising antecedents are associated with longer dependency lengths, an effect that is not reducible to a simple frequency effect. Moreover, the effect mostly exists within Indo-European and head-initial languages in our analysis, and is more consistent for subject relations. We discuss the cross-linguistic variability in General Discussion. In Study 2a and 2b, examining comprehension data from English reading-time corpora, we find a reduced locality effect at the retrieval site for more surprising antecedents. Consistent with Study 1, this effect is more pronounced in subject relations and is observed more reliably in the self-paced reading corpus (Study 2a) than in the eye-tracking corpus (Study 2b).

Study 1: Production side

We first examine strategic resource allocation in dependency locality in production. We predict that in production, the pressure to minimize dependency length can be relaxed when the antecedent contains novel and unexpected information. Consider the subject–verb dependency in the sentences below in (5):

(5) a. The evil *monster* in the tower *approached*...
b. The cute *monster* who stayed in the tower near the castle *approached*...

The subject "the cute monster" in (5b) is more surprising compared to "the evil monster" in (5a). According to our hypothesis, the unpredictable "cute monster" should be prioritized with more memory resources for encoding, and therefore is more capable of resisting the interference or decay introduced by the intervening material before the verb. As a result, compared to (5a), the less predictable antecedent in (5b) is able to tolerate more intervening material before being re-accessed at the retrieval site (i.e., the right codependent), leading to longer dependency length. We measure the predictability of word w at position t as surprisal S_t:

$$S_t \equiv -\log p\left(w_t \mid w_{<t}\right), \tag{3}$$

which is the negative log likelihood of the word w_t given its preceding context $w_{<t}$. The higher the surprisal, the less predictable a word is. Therefore, we predict a positive correlation between ANTECEDENT SURPRISAL and dependency length L in production data.

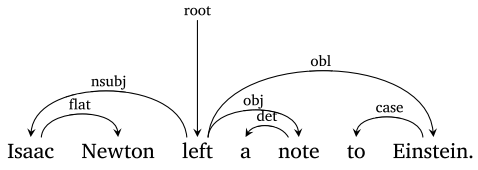
Besides ANTECEDENT SURPRISAL, we also examined the role of ANTECEDENT FREQUENCY in shaping memory allocation. On the one hand, these two quantities are highly correlated, in that low frequency words are also unpredictable in general, thus yielding higher surprisal. However, on the other hand, compared to surprisal, frequency as a unigram probability does not contain any information from the context. By comparing the effects of ANTECEDENT SURPRISAL and ANTECEDENT FREQUENCY, we aim to look into to what extent the contextual information contributes to the efficiency strategy of working memory encoding. We expect less frequent antecedents to associate with longer dependency length.

Method

Data

We used the corpora of 10 languages taken from Universal Dependencies (UD) release 2.11 (Nivre et al., 2020), as summarized in Table 1, with the aim to cover a wide variety of typological configurations (e.g., head-initial vs. head-final; free vs. rigid word order).[8] An illustration of UD annotations is shown in (6), where each arc represents a dependency whose direction is from the head to the dependent.[9]

(6) Example of UD annotation:



Compared to the Surface Syntactic Universal Dependencies (SUD) (Gerdes, Guillaume, Kahane, & Perrier, 2018), which is another major project of dependency corpora, the UD annotation scheme is content-word-oriented. That is, UD always labels content words as the head of a unit. As a result, UD favors lexical heads rather than functional heads in cases like adpositions, subordinating conjunctions, auxiliaries, and copulas. For the sentence above in (6), for example, SUD annotates the oblique relation as from the head "left" to the preposition "to" rather than to "Einstein". In the current work, we chose to use UD corpora since memory processes are more sensitive to content words rather than function words (Gibson, 1998; Grodner & Gibson, 2005).

Some UD corpora consist of out-of-context independent sentences, while others are organized document by document, which provide longer and enriched discourse context for each token. This difference may influence our surprisal estimates, which are sensitive to the preceding context of each token. We extracted all the dependencies of a sentence annotated in the UD corpora. All the UD corpora we used have a pre-defined split into training, dev, and test sets. We only used the pre-defined training sets since they already have decent sample size.

Estimating token surprisal

In this work, to ensure that the results are not the artifact of a specific language model, we generated surprisal measures from both the GPT-3 base (text-davinci-001 Brown et al., 2020) and the mGPT language models (Shliazhko et al., 2024), both being trained on multilingual data. For each token w in the dependency corpora, we obtained its surprisal $-\log p(w_t \mid w_{<t})$ given the preceding context from both models. We used the maximally allowed context window in the corresponding document or sentence. It is worth noting that Mandarin Chinese, unfortunately, is not supported by mGPT. Therefore, we only report the results with GPT-3 surprisal for Mandarin.

Contemporary large language models (LLMs) implemented with artificial neural networks provide state-of-the-art probabilistic measures of linguistic sequences and next-word predictions for the approximation of human predictive processing in psycholinguistics research (Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Wilcox, Gauthier, Hu, Qian, & Levy, 2020; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023; Xu, Chon, Liu, & Futrell, 2023). Empirically, the surprisal generated from LLMs highly correlates with human language processing difficulty indexed by both behavioral and neural responses (Goodkind & Bicknell, 2018; Hao, Mendelsohn, Sterneck, Martinez, & Frank, 2020; Hoover et al., 2023; Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Li & Ettinger, 2023; Schrimpf et al., 2021; Shain et al., 2024; Wilcox et al., 2023; Xu et al., 2023).

Measuring dependency length

We did the analysis with two different measures of dependency length L. The first measure is an orthographic one L_{O}, which is the number of words between the codependents of a dependency. The second measure is an information-theoretic one L_{I}, which sums up the surprisal of all words between codependents from w_i to w_{i+N}:

$$\begin{aligned} L_{\mathrm{I}} &= -\log p(w_{i\ldots i+N} \mid w_{<i}) \\ &= -\sum_{j=i}^{i+N} \log p(w_j \mid w_{<j}). \end{aligned} \tag{4}$$

We used these two measures because different words presumably induce memory interference to different extents. For example, compared to a content word that marks a discourse referent, a function word such as a determiner is way less informative, and may require much smaller memory load, thus inducing weaker memory interference (Gibson, 1998; Grodner & Gibson, 2005). Compared to the orthographic L_{O}, which treats all the words in the same way, the information-theoretic L_{I} may better capture the above-mentioned variability across different words (Hahn et al., 2021).[10]

Data transformation and exclusion

Constructions such as foreign phrases, multi-word proper names, and fixed expressions are annotated as flat structures in UD corpora. We merged flat structures such that the surprisal of the whole structure is the sum of all its components, and that the first word in the flat structure is treated as the head when calculating the length of a dependency. For example, the subject–verb dependency in (6) involves a flat structure in the subject position. The antecedent surprisal for this dependency is thus the sum of surprisal over both words "Isaac Newton", and the dependency length by word counts is 1, since the first word "Isaac" is one word away from the verb. We excluded sentences that are less than five-word long, since sentences that are too short may have limited room for the dependency length to vary and many of the short "sentences" are in fact titles and extended proper names (e.g., e-mail addresses and institution names). We excluded punctuation tokens. We also excluded tokens whose surprisal value is greater than 20 bits, as the surprisal estimates for such rare word sequences may be unreliable. Moreover, exceedingly surprising information may introduce confounding factors in human processing. We then extracted all the dependencies in which both the head and the dependent are spared from data exclusion.

[8] The original Russian corpus has over 1.2M tokens with over 600 documents; we randomly sampled 300 documents from the original corpus in our analysis in order to save on computational power.

[9] In our analysis, antecedent is defined as the left codependent of a dependency, and the retrieval site is always considered the right codependent, although as seen in (6) the direction of a dependency can either go from the left codependent to the right or the other way around.

[10] A potential problem with the information-theoretic L_{I} lies in the dual role attributed to surprisal: it has been theorized as being proportional both to the allocated memory resources and to memory cost. Although intuitively more memory resources allocated may induce higher memory cost as well, we acknowledge that the extent to which these two concepts can be treated as interchangeable remains a debatable question.

Table 1
Dependency corpora used in Study 1. 'Genre' refers to whether the texts in the corpus are organized as independent sentences ('sent'), or as documents with larger coherent discourse size ('doc'). '# All' indicates the number of all the dependencies after data exclusion. '# Subj' is a subset of '# All' and indicates the number of dependencies with subject relations. '# Obj' indicates the number of dependencies with object relations.

Language	Corpus	Genre	# All	# Subj	# Obj
Danish	DDT (Johannsen, Alonso, & Plank, 2015)	sent	45,976	4,203	3,963
English	GUM (Zeldes, 2017)	doc	89,947	7,881	7,296
German	GSD (McDonald et al., 2013)	sent	155,480	9,602	8,474
Italian	ISDT (Bosco, Montemagni, & Simi, 2013)	doc	208,939	10,323	11,735
Japanese	GSD (Tanaka et al., 2016)	sent	113,771	5,005	4,018
Korean	Kaist (Chun, Han, Hwang, & Choi, 2018)	doc	154,609	9,855	24,690
Mandarin	GSDSimp (Nivre et al., 2020)	sent	63,456	5,538	7,576
Russian	SynTagRus (Droganova, Lyashevskaya, & Zeman, 2018)	doc	329,745	32,822	25,065
Spanish	AnCora (Taulé, Martí, & Recasens, 2008)	doc	333,728	21,472	31,143
Turkish	BOUN (Marşan, Akkurt, Şen, Gürbüz, Güngör, Özateş, Üsküdarlı, Özgür, Güngör, & Öztürk, 2022)	sent	45,914	3,861	4,680

Data analysis

For each language, the analysis consists of three parts. The first one is on the full dataset obtained as introduced above, with all types of dependency relations included. In addition, we also took a closer look into the dependencies whose dependent is a core argument in the sentence. Therefore, we also ran analysis on two subsets of the full dataset above, which include subject relations[11] and object relations[12] respectively.

For the analysis with the full dataset, for each language, we ran separate linear mixed-effects models predicting the two variants of dependency length L as the dependent variable, using the `lmerTest` package in R (Kuznetsova, Brockhoff, & Christensen, 2017). The critical fixed-effect predictor is the ANTECEDENT SURPRISAL, with random intercept by dependency types.[13] For the analyses with subject and object relations, we ran linear models with the same fixed effects. We included five control variables for all the analyses, as in (7). SENTENCE POSITION aims to control the discourse-level information structure, where more information may be given as the discourse develops; ANTECEDENT POSITION aims to control that antecedents appearing towards to the end of a sentence naturally tend to have shorter dependency length. SENTENCE LENGTH aims to control for two possible confounds: first, longer sentences may tend to have longer dependency length in general; second, longer sentences may tend to have more complex syntactic structure, which may be associated with more surprising antecedents. We also included ANTECEDENT FREQUENCY in log scale retrieved from (Speer, 2022) in order to see whether the surprisal effect is reducible to a simple frequency effect. For the analysis with information-theoretic dependency length L_I, we included an additional control variable BASELINE SURPRISAL, which is the surprisal averaged across all words within a sentence. This is to address the confound that sentences with higher baseline surprisal naturally leads to a positive correlation between antecedent surprisal and the information-theoretic L_I. All variables are z-scaled.

(7) Control variables in Study 1

- SENTENCE POSITION: position of the sentence in the current document (*only included if the corpus is organized document-by-document*)
- ANTECEDENT POSITION: position of the antecedent in the current sentence
- SENTENCE LENGTH: length of the sentence measured as word counts
- ANTECEDENT FREQUENCY: log frequency of the left codependent
- BASELINE SURPRISAL: average surprisal across all words within a sentence (*only included for the analysis with information-theoretic dependency length* L_I)

Result

The result of the raw data in its original scale is presented in Fig. 5, which shows dependency length L as a function of ANTECEDENT SURPRISAL. The statistical result of the regression models is presented in Fig. 6 for the effects of ANTECEDENT SURPRISAL and ANTECEDENT FREQUENCY. A high-level summary of the statistical evidence across languages is shown in Fig. 7. Since we used the surprisal measure from GPT-3 and mGPT, we describe an effect as robust and independent of model parameterization if it is significant in the same direction both in the analysis with GPT-3 and in the one with mGPT (i.e., both positive or both negative, highlighted in dark red and dark blue in Fig. 7).[14] We describe the effect as partially confirmed and less robust if it reaches significance with only one of the language models (highlighted in light red and blue in Fig. 7). We describe the effect as inconclusive if it is not significant with any model, or if GPT-3 and mGPT show significantly conflicting result (i.e., significantly positive in one model but significantly negative in the other).[15]

All types of relations

Antecedent surprisal. In the analysis of the full dataset with all types of dependency relations, we indeed found a significant positive effect of ANTECEDENT SURPRISAL for six out of ten languages, whereby more surprising antecedents are associated with longer dependency length. Specifically, for both measures of dependency length L, there is a positive effect in Danish, English, German, Italian, Russian, and Spanish. However, for Japanese, Korean, Mandarin and Turkish, contrary to our prediction, there is a negative ANTECEDENT SURPRISAL effect.

[11] Annotated as `nsubj` and `csubj` in UD corpora.

[12] Annotated as `obj`, `iobj`, `ccomp`, and `xcomp` in UD corpora.

[13] As mentioned above, we also compare the effect of ANTECEDENT SURPRISAL with ANTECEDENT FREQUENCY in the current analysis. However, the models with random slopes for both effects rarely converge. Therefore, for better interpretability of the statistical result, we only included random intercept by dependency types.

[14] Since Mandarin is not available for mGPT, a critical effect is highlighted in dark red or blue in Fig. 7 even though we only have the result with GPT-3.

[15] The use of these terms (i.e., *robust*, *less robust*, *partially confirmed*, and *inconclusive*) is only for expository purpose, and does not imply any direct statistical robustness test.

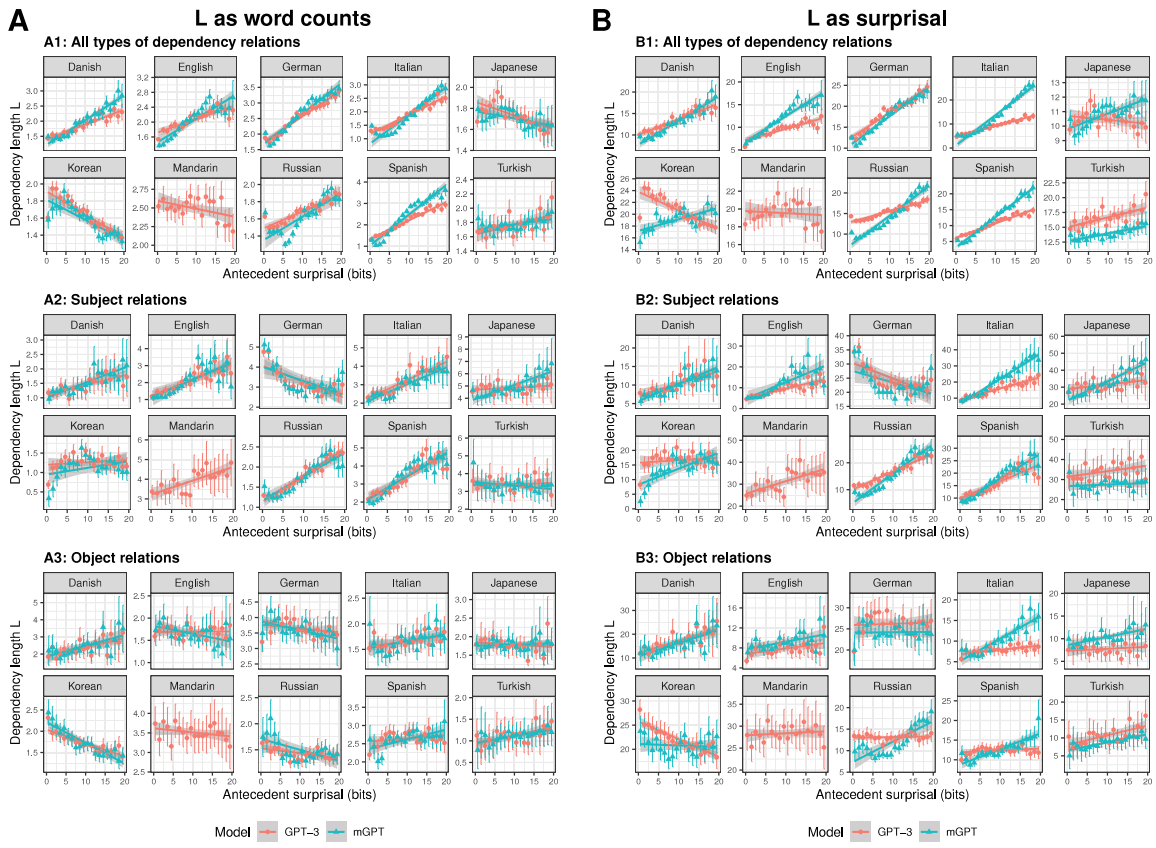


Fig. 5. Dependency length L as a function of ANTECEDENT SURPRISAL. ***Panel A*** corresponds to L measured as intervening word counts. ***Panel B*** corresponds to L measured as the sum of surprisal over intervening words. Surprisal is binned into 20 categories, and the mean L within each category is shown with a 95% confidence interval. A linear fit to these points is presented.

Antecedent frequency. The result is quite mixed for the effect of ANTECEDENT FREQUENCY. When L is measured as intervening word counts, there is a negative effect of ANTECEDENT FREQUENCY on L in six languages (robust for English, Italian, Russian, and Spanish, while partially supported in Korean and Turkish). The effect, however, is unexpectedly positive in German, Japanese, and Mandarin, and is inconclusive in Danish. When L is measured as surprisal, the ANTECEDENT FREQUENCY effect unexpectedly turns out to be positive for more languages, namely Danish, English, German, Japanese and Mandarin. There is still a negative effect for Russian, Spanish, Korean and Turkish.

Subject relations

Antecedent surprisal. For subject relations, when L is measured as word counts, there is a positive effect of ANTECEDENT SURPRISAL on L as predicted in five languages (English, Italian, Russian, Spanish, and Mandarin), suggesting that more surprising antecedents are associated with longer L in these languages. However, contrary to our prediction, the effect is negative for German, Japanese, Korean, and Turkish, while Danish shows inconclusive result. Similar pattern was observed when L is measured as intervening surprisal, except for Mandarin whose effect becomes inconclusive.

Antecedent frequency. The effect of ANTECEDENT FREQUENCY is the same for both measures of L. That is, there is a negative effect of ANTECEDENT FREQUENCY on L in six languages (Danish, English, Italian, Russian, Spanish, and Korean), suggesting that more frequent antecedents lead to shorter L in these languages. The effect is unexpectedly positive in German, and is inconclusive for Japanese, Mandarin, and Turkish.

Object relations

Antecedent surprisal. Surprisingly, for object relations, there is no positive effect of ANTECEDENT SURPRISAL on L in any language when L is measured as word counts. Instead, there is a negative effect in German, Russian, Spanish, Korean, and Mandarin, and the result is inconclusive for the rest of the languages. Similar pattern was observed when L is measured as surprisal, except that in Italian a positive effect is partially supported, and that the originally negative effect with orthographic L in Russian and Spanish becomes less robust.

Antecedent frequency. There is also a mixed picture for the effect of ANTECEDENT FREQUENCY in object relations. When L is measured as word counts, we only found a robust ANTECEDENT FREQUENCY effect on L in four languages, two negative (Italian and Spanish) and two positive (English and German). The result for the rest of the languages is inconclusive. When L is measured as intervening surprisal, there are three languages that show an unexpectedly positive effect (English, German, and Mandarin). Only in Spanish did we observe a negative ANTECEDENT FREQUENCY effect. The result for the rest of the languages remains inconclusive.

Discussion

In this cross-linguistic corpus study, we indeed found emerging evidence for a positive effect of antecedent surprisal on dependency length L, with both measures of L showing similar patterns. This effect still holds when we zoom into the subset that only includes subject or object relations. Overall, in many languages (especially Indo-Europeans), as predicted, this pattern indicates that more surprising antecedents are associated with longer dependency length, suggesting that the pressure to minimize dependency length is relaxed when the antecedent is of higher surprisal. Consistent with our hypothesis of strategic resource allocation, the result supports that novel and unexpected linguistic units can tolerate longer dependency length before its retrieval site,

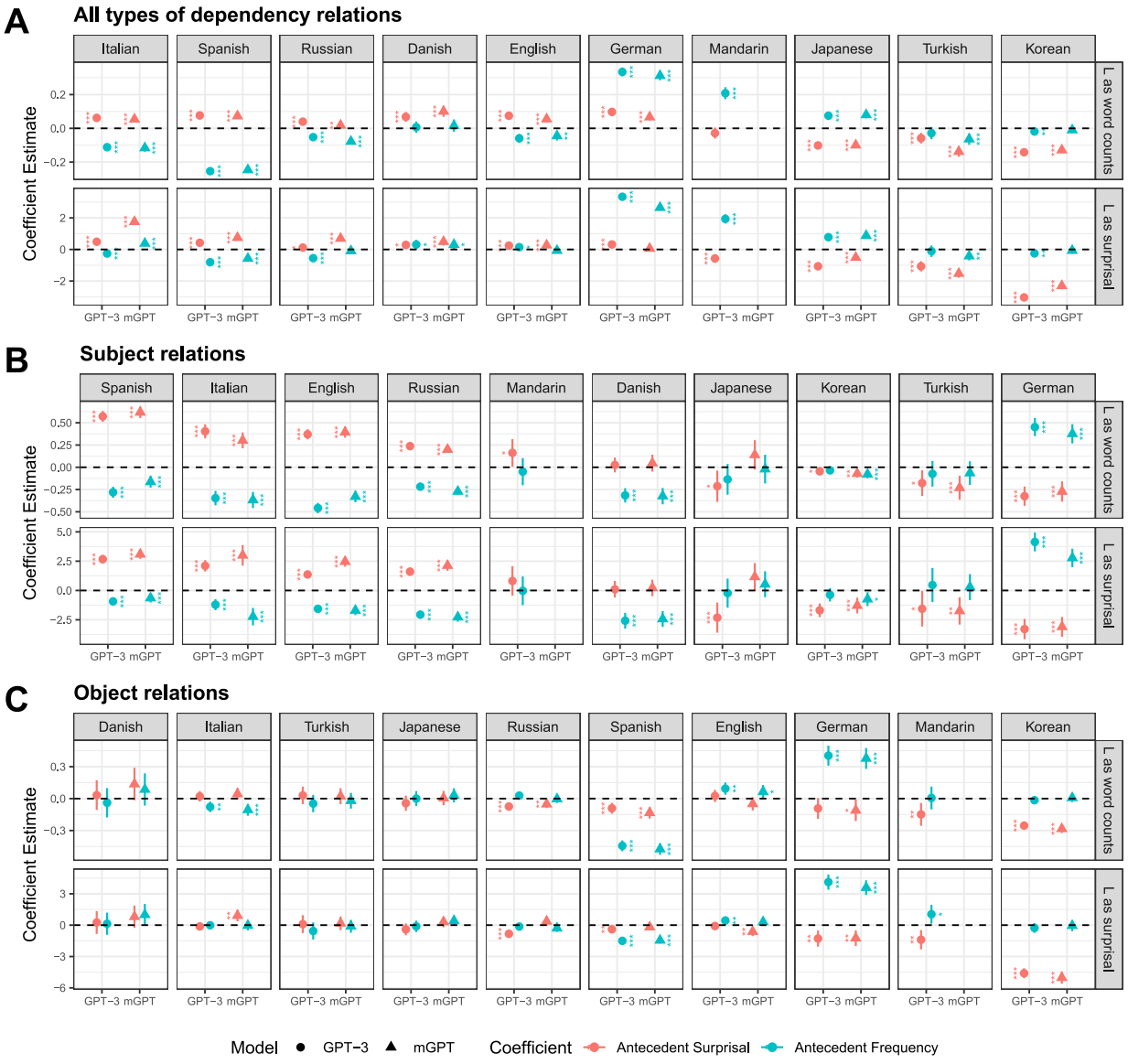


Fig. 6. Study 1 coefficient estimates for the effects of ANTECEDENT SURPRISAL and ANTECEDENT FREQUENCY on dependency length L across languages, with 95% confidence interval. Our hypothesis of strategic resource allocation predicts that more surprising antecedents are associated with longer L (positive effect of ANTECEDENT SURPRISAL), and that more frequent antecedents are associated with shorter L (negative effect of ANTECEDENT FREQUENCY). Significance levels: *(p<0.05), **(p<0.01), ***(p<0.001).

possibly because unexpected information is prioritized for working memory resources during encoding, and is more resistant to memory decay and interference.

However, there are two caveats worth noting. First, there is considerable cross-linguistic variability in our result, and the antecedent surprisal effect mostly exists within Indo-European and head-initial languages in our analysis. Second, although the analysis on the full dataset with all types of dependencies reveals a general trend for a positive antecedent surprisal effect, the result is much more consistent within subject relations. In object relations, the expected effect is reversed for most languages.

It is also worth noting that the positive ANTECEDENT SURPRISAL effect on L cannot be reduced to a pure frequency effect. That is, there is still a significant effect of ANTECEDENT SURPRISAL even though ANTECEDENT FREQUENCY has been included in the regression models as a control variable. Moreover, compared to ANTECEDENT SURPRISAL, the effect of ANTECEDENT FREQUENCY on L is less consistent.

In the end, to what extent does written corpus text approximate language production? Compared to spoken language, written language typically allows "speakers" more time to think, reducing much of the cognitive load involved in production, and the communicative goal is more geared towards listeners' need. That being said, speaker-oriented cognitive constraints, such as memory capacity, may play a less prominent role, and the need for strategic memory allocation may be diminished in written language production. Therefore, the effect observed in the current analysis using written text can be viewed as an lower bound, and we expect the effect of strategic memory allocation to be stronger when using spoken language corpora.

Study 2a: Comprehension side (Self-paced reading)

In this second study, we investigate whether the effect of strategic resource allocation also holds from the comprehension side. In particular, we examine to what extent the dependency locality effect observed

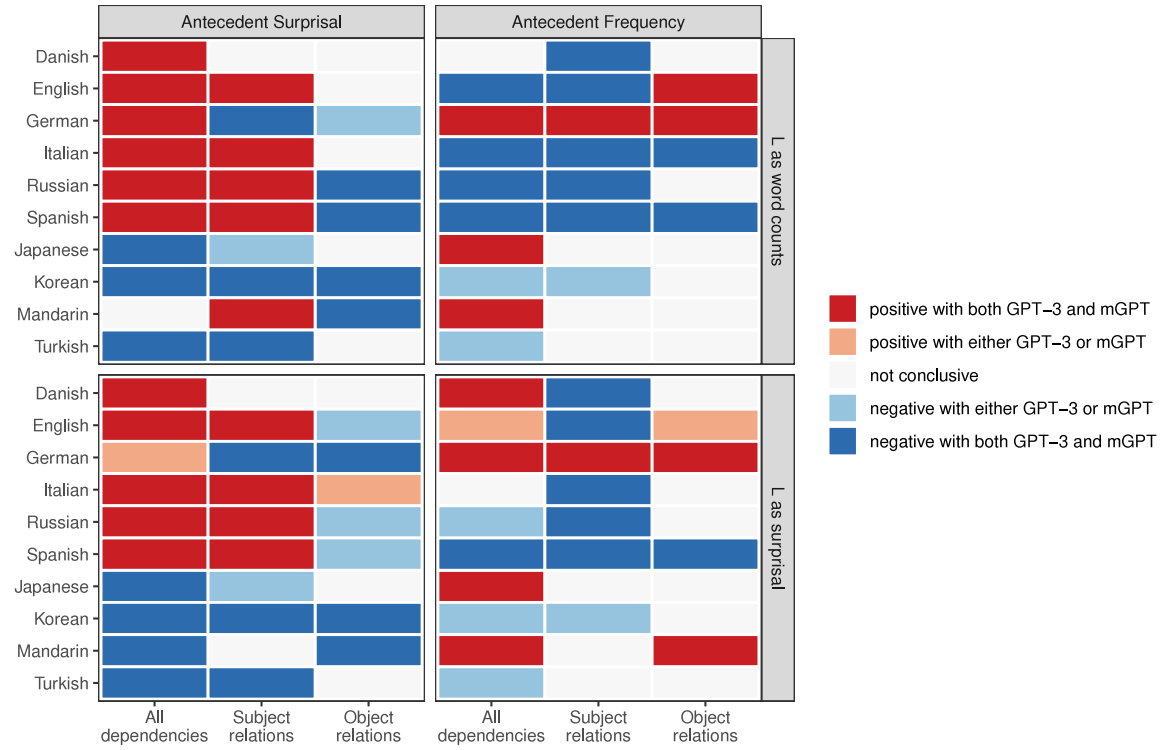


Fig. 7. Study 1 summary of statistical result for the effects of ANTECEDENT SURPRISAL and ANTECEDENT FREQUENCY on dependency length L. Significant ($p<0.05$) positive effects are highlighted in red; significant negative effects are highlighted in blue. Effects are considered not conclusive if insignificant with both language models, or if GPT-3 and mGPT show conflicting result where the effect is significant in opposite directions. According to our hypothesis, ANTECEDENT SURPRISAL is expected to have a positive effect on L, whereas ANTECEDENT FREQUENCY is expected to show negative effect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in previous comprehension studies (Bartek et al., 2011; Gibson, 1998, 2000; Grodner & Gibson, 2005) can be modulated by the surprisal of the antecedent, as illustrated in Fig. 8. First, we expect a baseline dependency locality effect, where the processing difficulty at the retrieval site, manifested as reading time, is expected to increase as the dependency length gets longer. Second, according to our hypothesis, more surprising antecedents are more capable of tolerating stronger memory interference. That is, the longer dependency length does not create too much additional processing difficulty for the retrieval of surprising antecedents, resulting in a reduced locality effect.

Consider the example in Fig. 8. As explained in Study 1, the more surprising "cute monster" in Fig. 8 (bottom) is more prioritized with encoding resources for its lower predictability, making it capable of tolerating longer dependency length. Therefore, when there is longer dependency distance, the processing difficulty at the retrieval site (i.e., the main verb "approached" in this example) should increase at slower rate for high surprisal antecedents than for low surprisal ones (Fig. 8; top), since the additional intervening material induces lower level of interference for more surprising antecedents. As a result, on top of the baseline dependency length effect on retrieval difficulty, we expect a negative interaction between dependency length L and ANTECEDENT SURPRISAL at the retrieval site.

Method

Data

The data we used in Study 2a is taken from the Natural Stories Corpus (NSC) (Futrell et al., 2021). The text of the corpus is in English, and contains 10,245 lexical words in 485 sentences, taken from 10 stories with around 1000 words each. The reading time (RT) data was collected from 181 native English speakers, using the self-paced reading task (SPR). The original corpus already excluded participants with low comprehension accuracy, as well as the reading times either faster than 100 ms or slower than 3000 ms. Therefore, we did not perform additional exclusion of reading time data in the current study. We generated the surprisal estimates for each word from mGPT.[16] The NSC corpus comes with dependency annotation in UD style.

Data transformation, exclusion, and analysis

As in Study 1, we analyzed three RT datasets as well here in the current study, namely the full dataset with all types of dependencies, a subset with subject relations only, and a subset with object relations only. The sample size is summarized in Table 2. We ran linear mixed-effect models on the log-transformed RTs of two regions, the critical region and the spillover region. The critical region is the right codependent of each dependency, which is considered the retrieval site for the antecedent. The spillover region is the word that goes immediately after the critical region. The critical effect is the interaction between dependency length L and ANTECEDENT SURPRISAL, with maximal converging random intercept and random slopes by participant. For the analysis of the full dataset, we also included maximal converging random effects by dependency type.

The control variables applied in Study 1 (namely, SENTENCE POSITION, ANTECEDENT POSITION, SENTENCE LENGTH, and ANTECEDENT FREQUENCY) are also included here in Study 2a. Besides these, we included several additional control variables that are often considered highly relevant for reading time measures. First, we included WORD LENGTH, WORD SURPRISAL and WORD FREQUENCY of the right codependent itself. Second, to control the spillover effect often found in reading studies, we included WORD SURPRISAL and WORD FREQUENCY of the two previous words before the right

[16] GPT-3 is no longer accessible from OpenAI since January 4th, 2024. Therefore, we only used the surprisal estimates from mGPT for the analyses in Study 2a and 2b.

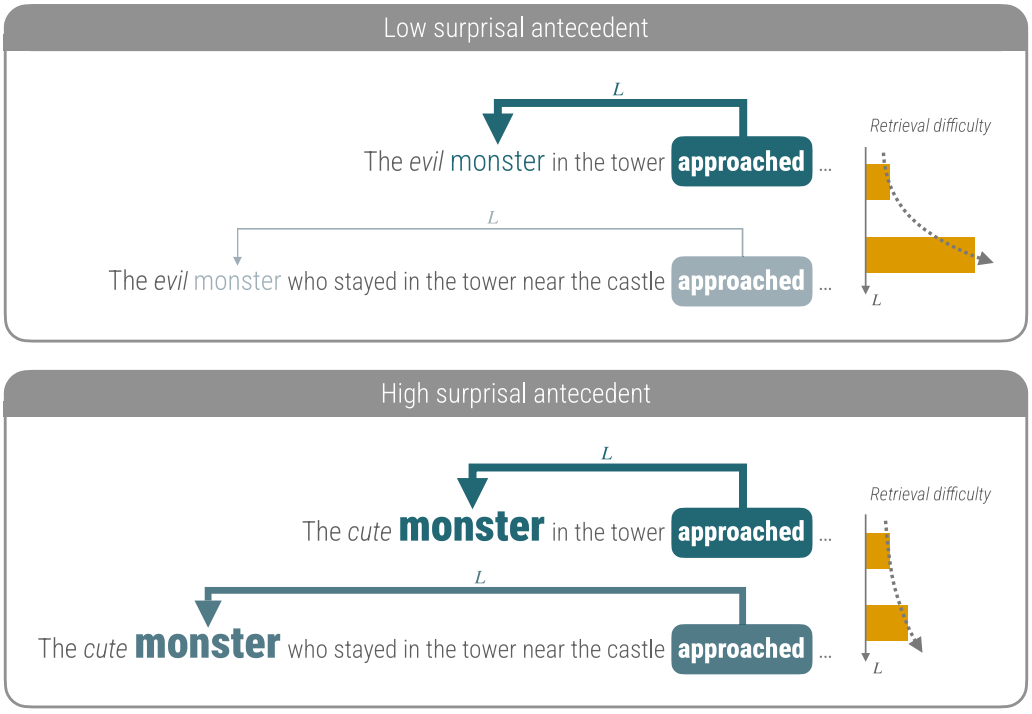


Fig. 8. Empirical prediction of strategic resource allocation in comprehension. Longer dependency length leads to higher processing difficulty at the retrieval site, and this retrieval difficulty increases more slowly for high-surprisal antecedents.

Table 2
RT data sample size in Study 2a and 2b.

	Study 2a	Study 2b	
		First-pass	Total RT
All types	601,122	256,567	313,167
Subject relations	57,023	21,709	26,885
Object relations	48,091	20,361	24,990

codependent. As in Study 1, frequency measures are in log scale, generated from (Speer, 2022). All variables are z-scaled. The transformation and exclusion of the dependency data follow the same procedure as in Study 1.

Results

Fig. 9A shows the interaction effect between dependency length L and ANTECEDENT SURPRISAL on raw reading times. The result of statistical models for the critical effects is summarized in Fig. 9B.

Critical region

All types of relations. In the analysis of the full dataset with all types of dependency relations, we found a baseline locality effect where dependency length L has a positive main effect on RTs at the retrieval site (i.e., the right codependent), suggesting that longer distance between codependents makes it more difficult to retrieve the antecedent at the right codependent. This main effect of L is only significant when L is measured as intervening surprisal. However, there is no significant main effect of ANTECEDENT SURPRISAL. Importantly, we found a negative $L \times$ ANTECEDENT SURPRISAL two-way interaction for both L measures. Consistent with our prediction, this negative interaction suggests that the locality effect on the RT of right codependents is reduced when the antecedent is more surprising.

Subject relations. For subject relation, although the main effect of L is numerically positive in the critical region, it is not significant with any measures of L. There is no ANTECEDENT SURPRISAL main effect, either. However, there is indeed a significant negative $L \times$ ANTECEDENT SURPRISAL two-way interaction for both L measures, suggesting a reduced locality effect for high surprisal antecedents.

Object relations. Again, for object relations, there is no main effect of L or ANTECEDENT SURPRISAL for any measures of L in the critical region. It is also worth noting that the L main effect is numerically negative, pointing to an anti-locality effect, although this effect is not statistically significant. Surprisingly, there is a positive $L \times$ ANTECEDENT SURPRISAL two-way interaction, although this effect is only significant when L is measured as word counts.

Spillover region

All types of relations. In the spillover region, we first found a baseline locality effect, where dependency length L leads to longer RT. This baseline locality effect holds for both measures of L. However, similar to the critical region, there is no evidence for an ANTECEDENT SURPRISAL main effect. In the end, again similar to the critical region, we found a negative $L \times$ ANTECEDENT SURPRISAL two-way interaction, suggesting that the locality effect is reduced when the antecedent is more surprising. This two-way interaction, however, only holds when L is measured as surprisal.

Subject relations. First, unlike the critical region, a baseline locality main effect of L was found for subject relations in the critical region, where longer L leads to longer RT at the right codependent. This L main effect is significant for both measures of dependency length L. Second, as in the critical region, there is no ANTECEDENT SURPRISAL main effect with either measure of L in the spillover region. In the end, we found a critical negative $L \times$ ANTECEDENT SURPRISAL two-way interaction with both L measures. As predicted, this negative interaction is indicative of a reduced locality effect for more surprising antecedents.

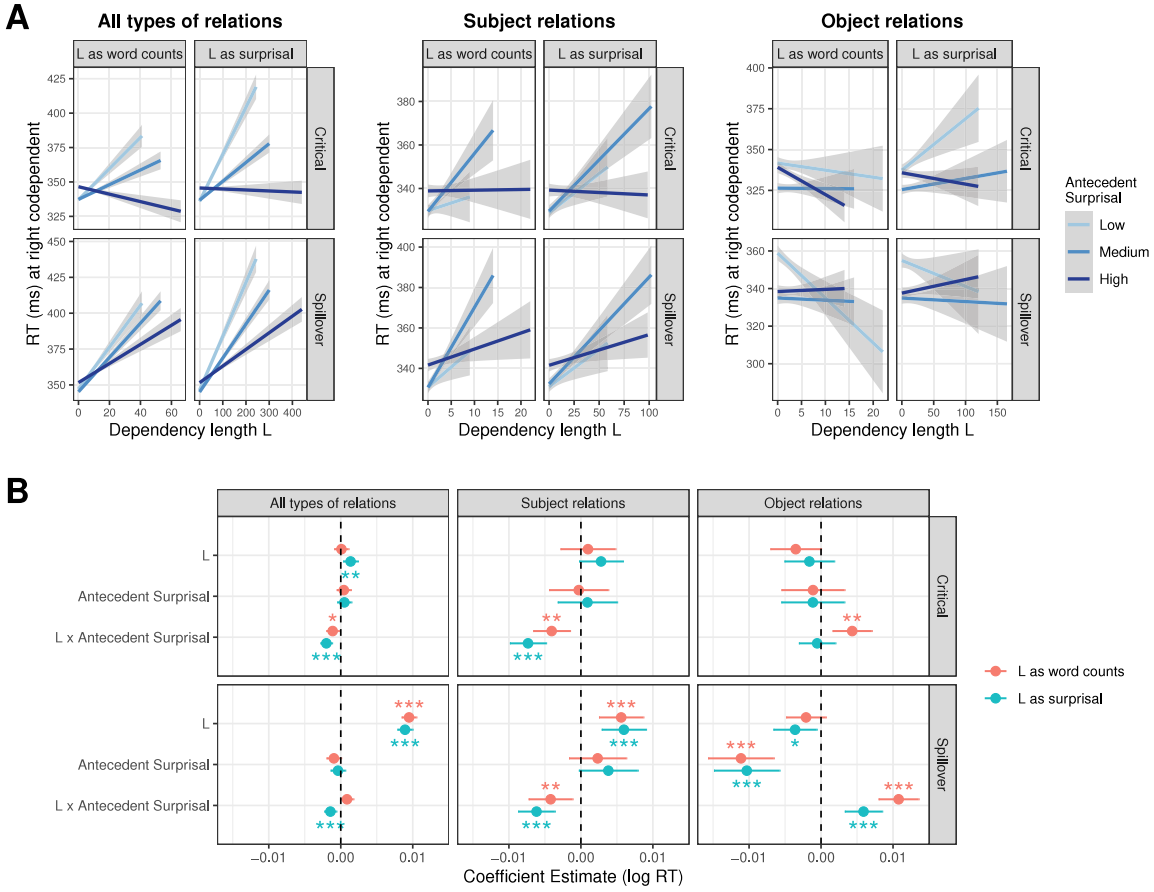


Fig. 9. Study 2a reading time (RT) result. ***Panel A***: raw RTs at the right codependent and its spillover region as a function of dependency length *L* modulated by ANTECEDENT SURPRISAL, which is binned into tertiles for visualization. ***Panel B***: result of regression models with log-transformed RTs; coefficient estimates with 95% confidence interval for the effects of *L*, ANTECEDENT SURPRISAL, and the interaction between the two. Significance levels: *(p<0.05), **(p<0.01), ***(p<0.001).

Object relations. The result of object relations has a relatively complex pattern. First, there is an unexpected negative, not positive, main effect of *L* on the RT at retrieval site. This negative main effect reaches significance when *L* is measured as surprisal, and still numerically holds when *L* is measured as word counts. Instead of a baseline locality effect, this negative *L* effect suggests an *anti*-locality effect, where more intervening material between the two codependents leads to faster RT at the retrieval site. Moreover, as seen in Fig. 9A (right column), this *anti*-locality effect is mostly driven by antecedents of low surprisal. Second, there is a negative ANTECEDENT SURPRISAL main effect for both *L* measures, whereby more surprising antecedents induce faster RT at the retrieval site. In the end, unlike the negative *L* × ANTECEDENT SURPRISAL interaction observed in the previous two analyses, object relations exhibit a positive *L* × ANTECEDENT SURPRISAL interaction, which is significant for both *L* measures. However, since the main effect of *L* is negative in the first place, instead of an enhanced locality effect, this positive interaction actually suggests a reduced *anti*-locality effect for more surprising antecedents. It is not yet entirely clear to us why there is an *anti*-locality effect in the first place exclusively for object relations, but the result pattern in object relations seems to point to a potential trade-off between the direction of the *L* main effect and the direction of the *L* × ANTECEDENT SURPRISAL interaction, which we will discuss below.

Discussion

To sum up, in Study 2a, through the analysis of the RT data in the Natural Stories Corpus, we first replicated the baseline locality effect in the analysis of all types of dependency relations and subject relations, especially in the spillover region. This suggests that the non-local retrieval of the antecedent at the right codependent becomes more difficult when there is more intervening material. However, we also found an *anti*-locality effect in object relations in the spillover region, which suggests that more intervening material actually facilitates the establishment of a non-local object relation. Second, only in object relations did we observe a negative main effect of ANTECEDENT SURPRISAL on the RT at the right codependent, but this effect only emerges in the spillover region. This negative ANTECEDENT SURPRISAL main effect suggests that more surprising antecedents are easier to retrieve in object relations, which is consistent with the findings in Hofmeister (2011), where semantically more complex noun phrases are easier to be retrieved. However, more future work is needed to investigate why this effect only emerges in object relations. In the end, the most important finding of this Study 2a is the interaction between dependency length and antecedent surprisal, both in the critical and the spillover regions. Specifically, in the analysis of the full dataset and the one of subject relations, we found a reduced locality effect for more surprising antecedents. This reduced locality effect is consistent with the prediction of strategic resource allocation, in that more surprising antecedents are prioritized for working memory resources and are encoded with more robust representation against memory interference and decay.

It is also worth noting that there seems to be a potential trade-off between the direction of the *L* main effect and the direction of the *L* × ANTECEDENT SURPRISAL interaction. On the one hand, when the main effect of *L* is positive, there is a negative *L* × ANTECEDENT SURPRISAL interaction, suggesting a reduced locality effect for more surprising antecedents. On the other hand, when the main effect of *L* is negative to

start with, as in the object relations, the interaction becomes positive, indicating that the *anti*-locality effect is reduced for more surprising antecedents. The reduced locality effect is straightforward, as predicted by our hypothesis. But why is there a reduced *anti*-locality effect? In fact, the reduced *anti*-locality effect for more surprising antecedents can be consistent with our strategic resource allocation as well. According to experience-based processing theories, the *anti*-locality effect can be viewed as a facilitation effect on the prediction of the right codependent. That is, more intervening material may provide more information about the identity of the word at the right codependent, helping the comprehender to make better predictions (Levy, 2008a), canceling out the burden created by memory interference. However, for more surprising antecedents, if their representation is more enhanced due to strategic resource allocation, it is possible that comprehenders can already rely on the their memory of the antecedent to predict the right codependent. As a result, the intervening material may no longer provide too much additional help to make predictions. This reduced facilitation from the intervening material for more surprising antecedent, therefore, may manifest itself as a reduced *anti*-locality effect.

Study 2b: Comprehension side (Eye-tracking)

In Study 2b, we examine strategic resource allocation in dependency locality using an eye-tracking corpus. As in Study 2a, we expect to see a baseline locality effect, as well as an interaction between locality and antecedent surprisal, in the sense that a reduced locality effect is associated with more surprising antecedents.

Method

Data

The data we used in Study 2b is taken from the English part of Dundee corpus (Kennedy & Pynte, 2005). The corpus consists of 20 texts, with 56,212 tokens in total (around 2800 words for each text). The eye-tracking data is collected from 10 English native speakers, with each text being split into 40 fine-line screens. We analyzed two eye-tracking measures: *first-pass reading time*, defined as the sum of all the fixations on a region after first entering in the region and before first leaving it either to the left or to the right; and *total reading time*, defined as the sum of all the fixations on a region throughout a trial. Like NSC corpus, the Dundee corpus also comes with UD-style dependency annotation.

Data transformation, exclusion, and analysis

The transformation and exclusion of dependency data follow the same procedure as in Study 1 and Study 2a. The reading time responses are excluded if shorter than 100 ms or longer than 3000 ms. The sample size after data exclusion is summarized in Table 2. We ran linear mixed-effects models on first-pass durations and total reading times. Both reading time measures are log-transformed. As in Study 2a, the critical effect is the interaction between dependency length L and ANTECEDENT SURPRISAL, with the same random structure and control variables as in Study 2a.

Results

Fig. 10A shows the interaction effect between dependency length L and ANTECEDENT SURPRISAL on raw reading times, including both the first-pass duration and the total RT. The result of statistical models for the critical effects is summarized in Fig. 10B.[17]

All types of relations. As shown in Fig. 10B (left column), no evidence was found for the main effect of dependency length L with any measure of RT and L, suggesting the lack of the baseline locality effect. No significant main effect of ANTECEDENT SURPRISAL was found, either. For the critical $L \times$ ANTECEDENT SURPRISAL interaction, although the effect is numerically negative on first-pass RT with both measures of L, it is only marginally significant. No evidence for the interaction effect was found on total RT.

Subject relations. Still, as shown in Fig. 10B (middle column), we did not find any evidence for either an L or an ANTECEDENT SURPRISAL main effect with any measure of RT and L. However, there is indeed a negative $L \times$ ANTECEDENT SURPRISAL two-way interaction, suggesting that the locality effect at the retrieval site, although not statistically significant on average, is reduced for more surprising antecedents. This interaction effect reliably holds for first-pass RT with both L measures, as well as for total RT with L measured as word counts. It is, however, only marginally significant for total RT with L as intervening surprisal.

Object relations. As shown in Fig. 10B (right column), we found in object relations a baseline locality effect manifested as a positive L main effect on RTs at the retrieval site, which holds for first-pass RT with L as word counts and for total RT with both L measures. There is no evidence for an ANTECEDENT SURPRISAL main effect. In terms of the critical $L \times$ ANTECEDENT SURPRISAL interaction, we only found a marginally significant negative interaction for total RT with L measured as surprisal, and the effect is non-significant elsewhere.

Discussion

The result of the two main effects in the current Study 2b shows a very different pattern compared to Study 2a. First, unlike Study 2a, in the current Study 2b only in object relations did we find a baseline locality effect, whereas the locality effect is not observed in the analysis of the full dataset or in the one of subject relations. That is, longer dependency length does not lead to higher processing difficulty at the retrieval site for subject relations. This lack of the baseline locality effect aligns with the observation in Demberg and Keller (2008), where the locality effect in Dundee corpus is overall small and unreliable for verbs, which in our case is the retrieval site of subject relations. Second, there is no main effect of ANTECEDENT SURPRISAL across the board in this Study 2b.

Although the baseline locality effect is relatively unreliable, we still observed evidence for a negative $L \times$ ANTECEDENT SURPRISAL interaction effect in this Study 2b, especially in subject relations. As shown in Fig. 10A (middle column), compared to antecedents with low-to-mid surprisal levels, those with high surprisal exhibit weaker locality effect. Similar to the interaction effect observed in Study 2a, the current result shows that more surprising antecedents is less susceptible to the effect of memory interference induced by intervening material, possibly due to their enhanced representation. Supplementing the self-paced reading data in Study 2a, the result of Study 2b thus lends support to our hypothesis of strategic resource allocation with data from eye-tracking paradigm.

General discussion

In three corpus studies, we examined strategic resource allocation (SRA) through the lens of dependency locality both in production and in comprehension. Study 1 explored this hypothesis in production by analyzing UD corpora of 10 languages. Our result reveals that more surprising antecedents can tolerate more intervening material before they need to be retrieved at the other side of the dependency, resulting in a positive correlation between antecedent surprisal and dependency length. However, it is worth noting that this reduced locality effect mostly exists within Indo-European and head-initial languages, and is

[17] No critical effects were found in the spillover region, so we only report the result of the critical region in this Study 2b.

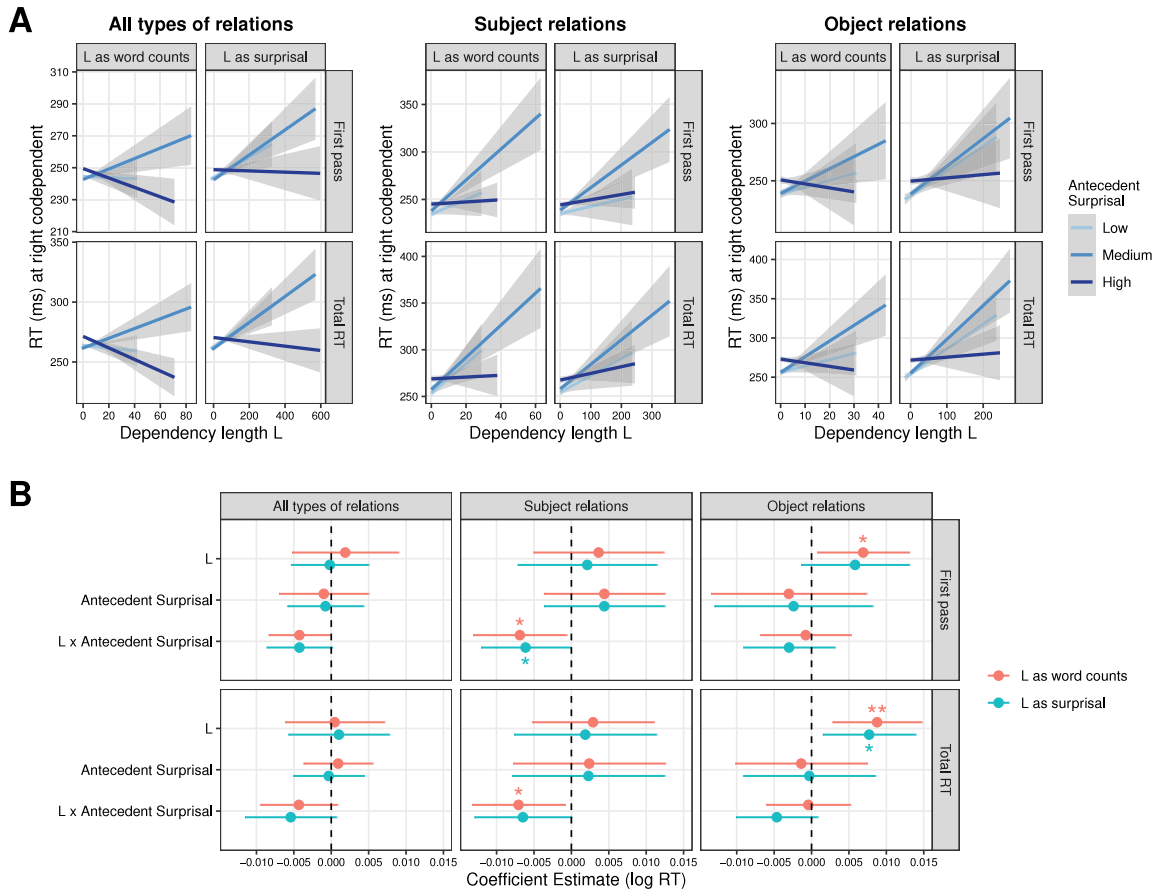


Fig. 10. Study 2b reading time (RT) result for first-pass duration and total RT. ***Panel A***: raw RTs at the right codependent as a function of dependency length L modulated by ANTECEDENT SURPRISAL, which is binned into tertiles for visualization. ***Panel B***: result of regression models with log-transformed RTs; coefficient estimates with 95% confidence interval for the effects of L, ANTECEDENT SURPRISAL, and the interaction between the two. Significance levels: *(p<0.05), **(p<0.01), ***(p<0.001).

more consistent for dependencies of subject relations than for object relations.

In Study 2, we shifted gears and focused on comprehension, analyzing two English reading-time corpora: one based on self-paced reading paradigm (Study 2a) and the other based on eye-tracking paradigm (Study 2b). The SPR data reveals a baseline locality effect, where longer dependency lengths lead to increased reading times at the retrieval site. Importantly, we found a $L \times$ ANTECEDENT SURPRISAL interaction, where the baseline locality effect is reduced for more surprising antecedents. Moreover, we again observed a subject–object asymmetry, such that the critical effect consistently holds only in subject relations. The eye-tracking data shows a more nuanced pattern: although the baseline locality effect was observed only in object relations, we indeed found an $L \times$ ANTECEDENT SURPRISAL interaction in subject relations.

Overall, despite the caveats mentioned above, our result shows emerging evidence that a reduced locality effect emerges for more surprising antecedents in the processing of non-local dependencies, suggesting that more surprising antecedents are less susceptible to the interference from intervening material. This finding aligns with the notion of strategic resource allocation that we proposed, which holds that unexpected information is prioritized for memory resources and is encoded with enhanced memory representation.

Processing difficulty as encoding difficulty: Reinterpreting the effect of surprisal and entropy

In this section, we first reinterpret the processing difficulty of a word as its memory encoding difficulty. As shown in Fig. 11, the encoding process in Fig. 2 can be considered a transformation from a flat uniform distribution over all possible words to one that is concentrated around the received input. The processing difficulty of a word at the encoding stage, therefore, can be considered the distance between the pre-encoding and the post-encoding distribution. As a result, less uncertain, or more precise encoding distribution is more distant away from the uniform pre-encoding distribution, leading to higher processing difficulty. With this encoding view in mind, let us consider two factors that have been previously argued to influence the processing difficulty of a word, namely surprisal and entropy.

For the surprisal effect, as shown in our main proposal of SRA, more surprising input should be encoded with higher precision, an efficiency strategy that we have argued to minimize the retrieval error at a later time point. As a result, SRA naturally predicts that the more precise encoding for more surprising inputs should lead to higher encoding or processing difficulty, which is consistent with the widely observed surprisal effect in the literature. Importantly, this memory encoding view provides a resource-rational account for the surprisal effect, reinterpreting it as a strategic solution to the efficiency problem of memory. We will discuss this in more detail below in Section "Surprisal effect as efficient coding: An adaptionist view".

For the effect of entropy, SRA yields complicated predictions. As mentioned in Introduction, SRA implies that higher uncertainty of prior does not necessarily increase or decrease the precision of encoding distribution, and therefore does not necessarily increase or decrease the processing difficulty of a word. To demonstrate the reason behind it, first recall that according to SRA the encoding precision depends on how the received input can be accurately reconstructed at a later

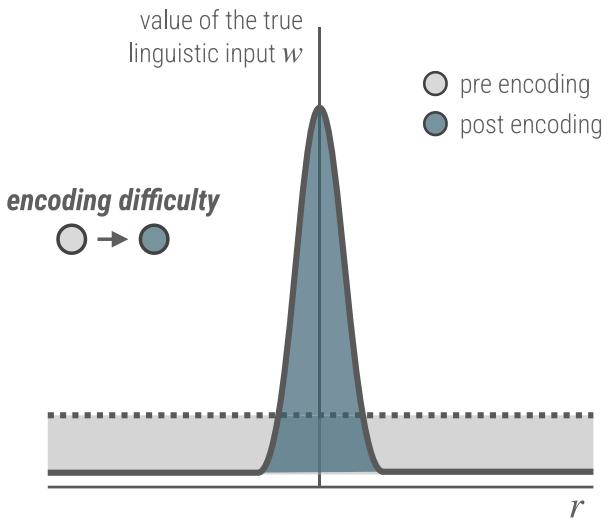


Fig. 11. Processing difficulty as memory encoding difficulty.

time point, which is in turn dependent on to what extent such a reconstruction can be supported by the prior. Intuitively, when the prior is highly consistent with the actual received word (e.g., low surprisal words), the more precise the prior is, the less precise the encoding of that word needs to be for better reconstruction. In this case, the uncertainty of prior should have a positive effect on encoding difficulty, in the sense that the processing of a word is facilitated if its preceding context yields a high-constraining prior. However, when the prior is not compatible with the actual input (e.g., high surprisal words), such a facilitation effect is necessarily the case any more: a highly precise but incompatible prior may actually need to be counteracted by more effort into precisely encode the actual input. In fact, this complicated effect of prior precision echos the empirical observation in many previous studies, where the effect of prediction entropy is much less reliable than the effect of surprisal on processing difficulty (Linzen & Jaeger, 2016; van Schijndel & Linzen, 2021; Wilcox et al., 2023).

The role of representational uncertainty

Under SRA, there is a dissociation between *accuracy* and *uncertainty* in memory representations. For accuracy, it is more related to the *mean*, or *point estimate* of an underlying distribution with respect to how far it is from the true input value. For uncertainty, it corresponds to the *precision*, or *variance* of the distribution, reflecting the relative competitiveness of all alternative inputs. As mentioned in Introduction, the critical prediction of SRA is about the *precision* or *uncertainty* in the encoded memory representation, as reflected in the robustness against interference, rather than about the raw *accuracy* of retrieval. As shown in Fig. 2, although a relatively high retrieval accuracy may be maintained on average for all linguistic inputs, those that receive more resources will have less uncertainty in their encoded representation.

Both the point estimate and the uncertainty are important information to understand the underlying representations of memory processes, as raised by more and more recent work in psychophysics (Bays et al., 2024). However, compared to point estimates, the characterization of representational uncertainty is relatively underexplored in psycholinguistics research, both theoretically and empirically. For example, studies under the framework of cue-based retrieval often concern what representation has actually been retrieved, based on which an empirical prediction is derived. Similarly, studies under the noisy-channel framework often focus more on interpreting the point estimate of posterior distribution, rather than how the probability mass is distributed over the hypothesis space.

One of the major challenges to understand the role of representational uncertainty is possibly the lack of a straightforward linking hypothesis. In most psycholinguistics studies, the interpretation of *online* dependent measures such as reading times is based on the point estimate of its mean, which is naturally linked to the point estimate of mental representations (cf. Huang & Dillon, 2023). But for representational uncertainty, how this psychological construct is linked to any online behavioral measure remains unclear. Even though there might be some ways to probe the degree of representational uncertainty through certain *offline* measures (e.g., tasks that directly probe the errors in the interpretation of a sentence), it is still challenging to lay out the hypothesis space of alternative representations in a fine-grained manner.

In the current study, an important assumption we made is that less uncertainty leads to more robust representation against interference. That is, if the processor is more uncertain about the representation of a linguistic input, its encoding distribution may be more likely to be influenced and distorted by other information it holds in memory. Surely, this is a debatable assumption, and the specific mechanism of how the representation is distorted by other inputs needs to be further elaborated in future work.

The role of context in working memory efficiency

Earlier in Introduction, we have argued that the efficiency of working memory allocation depends on how likely a linguistic unit can be reconstructed based on the statistical structure of linguistic input. But an open question is: what kinds of statistics are being used? More specifically, to what extent is working memory efficiency guided by context-sensitive statistics?

The role of contextual information has been under debate across multiple domains of linguistic efficiency, especially in the context of signal reduction (Jaeger & Buz, 2017). One of the earliest evidence supporting the communicative efficiency pressure for linguistic structure is the well-known Zipf's law, which observes that more frequent words tend to have short forms (Zipf, 1949). Similar reduction effect of linguistic forms as a function of usage frequency has also been found in the historical change of linguistic representations (e.g., Bybee, 2006; Bybee, Perkins, & Pagliuca, 1994; Cohen Priva, 2015; Pierrehumbert, 2008). While most of these theories focus on frequency, which can be considered unigram probabilities independently generated from a stationary distribution, recent studies have begun investigating the role of context-specific probabilities in linguistic efficiency. The findings in this area are mixed. For example, building on Zipf's observation, Piantadosi, Tily, and Gibson (2011) find that a significant amount of word-length variability is explained by contextual predictability in addition to the frequency effect. In contrast, Pimentel, Meister, Wilcox, Mahowald, and Cotterell (2023) argue that word length is better predicted by frequency. Beyond the structure of lexicons, in online processing, contextual predictability also shapes the reduction of referring expressions (e.g., Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Tily & Piantadosi, 2009; Xu & Xiang, 2021) as well as syntactic structure (e.g., Jaeger, 2010; Jaeger & Levy, 2006).

In the current study, that the observed antecedent surprisal effect cannot be reduced to a pure frequency effect suggests that the statistics relevant to working memory efficiency go beyond unigram frequencies, and the strategic allocation of working memory resources is based on more fine-grained context-specific probabilities. As noted by Jaeger and Buz (2017), frequency can be understood as an averaged effect of contextual predictability. This is probably one of the reasons why linguistic theories focusing on the representational aspects of language often emphasize frequency, as it reflects the abstract, global properties of a language accumulated through long-term experience. That being said, the optimization of working memory efficiency not only relies on those ready-to-retrieve statistics already stored in long-term memory, but also incorporates statistics computed online, dynamically drawing

upon rich contextual information in a rapid and adaptive manner (for a different view, see Opedal, Chodroff, Cotterell, & Wilcox, 2024).

Moreover, the role of context in SRA has its implications for the asymmetry between proactive and retroactive interference. Proactive interference refers to the configuration where distractor information precedes the retrieval target, whereas for retroactive interference the distractor is located between the retrieval target and the retrieval site. Previous studies have observed that proactive interference has weaker effect than the retroactive one (Van Dyke & McElree, 2011), an empirical pattern that supports time-based decay of memory activation (Barrouillet, Bernardin, & Camos, 2004; Lewis & Vasishth, 2005; Page & Norris, 1998; Portrat, Barrouillet, & Camos, 2008). Under SRA, this asymmetry can be potentially explained from a resource-rational perspective. As shown in Fig. 2, at the encoding stage of a linguistic unit, the processor already has access to the information in its preceding context, but not yet to the upcoming information in the right context. Therefore, during memory encoding, the strategic allocation is possibly based only on the preceding context. In other words, the encoded representation of an input is optimized for information it has already received in the preceding context, but not necessarily for what has not yet been received. If this is the case, SRA naturally explains why the distractor that goes before the target unit has less impact on its representation than the retroactive distractor, without necessarily resorting to a separate time-based decay mechanism.

Relationship with lossy-context surprisal

The theoretical framework of SRA we proposed shares similar theoretical and empirical implications with lossy-context surprisal theory and its variants (Futrell, Gibson, & Levy, 2020; Hahn et al., 2022). Focusing on the prediction mechanism, lossy-context surprisal holds that next-word predictions are based on lossy and faulty memory representations, rather than the veridical form of the past linguistic input. The theory explicitly includes a memory distortion process, where certain elements in an utterance are erased to form a lossy representation, subject to certain probabilistic erasure distributions.

In many aspects, our proposal and lossy-context surprisal form two sides of the same coin. For our SRA, we seek to understand and explain the working memory mechanism in sentence processing, and predictive processing is a component incorporated into the mechanism we proposed to better explain memory. For lossy-context surprisal, in contrast, the theory aims to understand the prediction mechanism in sentence processing, with a memory component included to better explain prediction. Despite these different goals, we both speak for an interaction between memory and prediction, as both of them should jointly support human linguistic behaviors as a cognitive task. That means, our proposal and lossy-context surprisal are not mutually exclusive, theoretically or empirically, and we simply focus on different perspectives of a similar cognitive task.

In fact, our theoretical framework of SRA is on some level mutually translatable with lossy-context surprisal, so we do not see them necessarily as conflicting theories. On the one hand, in the language of lossy-context surprisal, the memory distortion process where certain linguistic units are erased is potentially where our strategic resource allocation could fit in, such that more surprising units given the context are less likely to be erased when predicting future units. On the other hand, a fundamental question at the core of our resource-rational analysis of working memory mechanism is: if memory capacity is limited, how to minimize the cost of memory error by strategically prioritizing more important linguistic units? However, what counts as important? Or, in other words, what is the objective function based on which the cost of memory error is defined? In the original lossy-context surprisal theory (Futrell, Gibson, & Levy, 2020), this objective function is not explicitly specified. In Hahn et al. (2022), the model takes one step further, and this objective function is to minimize the downstream next-word prediction task. In our proposal, the cost is defined by how likely a unit can be reconstructed later given the context. These two different objective functions of cost are not mutually exclusive, and the accuracy of next-word prediction may be compromised if a lost unit in memory is not reconstructable. Therefore, although our proposal of SRA has a different theoretical focus from the model of Hahn et al. (2022), our empirical predictions actually share some overlap, and we both predict on some level that the representation of more surprising units (or, less frequent units) should be enhanced and more robust.

Hierarchical encoding and compression

SRA, arising as an efficiency principle from the functional pressures of working memory, is situated more at the computational level in Marr (1982)'s three-level representation. A natural question to ask then is: what is the potential mechanism to implement this efficiency principle at the algorithmic level? In other words, when more resources are allocated, what makes the representation less uncertain and more robust?

One possible mechanism is hierarchical compression, which postulates that information can be stored in memory with a multi-level hierarchy of abstraction (Bates & Jacobs, 2020; Brady, Konkle, & Alvarez, 2009; Brady et al., 2024; Christiansen & Chater, 2016; Craik & Lockhart, 1972). In visual working memory, higher levels are more compressed, having a more categorical nature; lower levels, in contrast, are encoded with more quantitative perceptual detail. A similar hierarchy also exists in sentence processing. Sequential linguistic input can be continuously encoded and recoded into compressed forms, which in turn are further compressed into more abstract forms when new input comes in Christiansen and Chater (2016). This incremental compression procedure gives rise to a multi-level hierarchical structure of memory representation, such that higher levels of abstraction are encoded as a gist of message without specifying elaborated syntactic and semantic features (Bradshaw & Anderson, 1982). Intuitively, more memory resources should yield more detailed encoding. Indeed, this has been recently demonstrated by some memory encoding models grounded in the rate–distortion theory, where a quantitative-categorical spectrum naturally arises simply by manipulating the memory capacity during encoding (Bates & Jacobs, 2020; Jakob & Gershman, 2023).

Predictable information, if less prioritized, should be encoded in a more compressed and abstract fashion. In visual working memory, this is indeed evidenced by the fact that more memory objects can be stored in visual working memory tasks when their perceptual features are statistically correlated (Bates, Lerch, Sims, & Jacobs, 2019; Brady et al., 2009). Moreover, information with stronger prior knowledge is also more susceptible to the categorical bias in perception, where the perceived input is biased towards the categorical mean (Bates & Jacobs, 2020). Similarly, in the domain of language, more frequent linguistic sequence is more likely to be holistically stored in memory (e.g., Bybee, 2006; Goldberg, 2003; Hawkins, 2004; Traugott & Trousdale, 2013). Importantly, hierarchical compression based on statistical regularities provides a mechanism that strings together effects across different linguistic representational levels, forming a spectrum of compression. At one end, there is the locality effect where words that are more mutually predictable tend to stay closer to each other in linear order (e.g., Futrell, 2019; Futrell, Qian, Gibson, Fedorenko, & Blank, 2019). At the other end, the same pressure of compression governs the fusion of morphemes (e.g., Hahn et al., 2021; Rathi, Hahn, & Futrell, 2021), and makes mutually predictable units more likely to go through processes such as affixation and phonological reduction (e.g., Bybee, 2006; Bybee et al., 1994; Gahl & Baayen, 2024).

Parallelism between production and comprehension

The effect of SRA holds both for production and for comprehension in the current study, pointing to a parallelism between these two modalities. For comprehension, a reasonable question is: to what

extent the observed effect of strategic resource allocation is experience-based (MacDonald & Christiansen, 2002), given that the same effect is also seen in production data? In other words, it is possible that comprehenders prioritize unexpected linguistic units and encode them with enhanced representation because unexpected units are more likely to associate with stronger memory interference in the production data they receive.

A similar question can be asked for production as well: to what extent is the effect observed in production the result of audience design (Clark & Murphy, 1982; Ferreira, 2019; Lockridge & Brennan, 2002), given that comprehenders can encode unexpected units with prioritized memory resources? One listener-oriented production theory compatible with our finding is the Uniform Information Density (UID) theory (Clark et al., 2023; Jaeger & Levy, 2006; Meister et al., 2021). According to UID, surprising antecedents may be followed by longer dependency length so that there is a smoother transition to the other side of the dependency, which is relatively predictable since it shares high mutual information with the antecedent (Futrell et al., 2019).

Here we do not attempt to adjudicate between the two questions above, nor do we view them in conflict with our proposal. In our opinion, SRA provides a potential explanation for the mechanistic underpinnings of these higher-level processes.

Despite the parallelism, comprehension and production still differ in some critical aspects, exerting modality-specific constraints on SRA due to their idiosyncratic processing nature. For example, production is, in general, more cognitively demanding, in need of higher memory capacity, executive control, and action planning than comprehension (Hickok, 2012; Koranda, Bulgarelli, Weiss, & MacDonald, 2020; MacDonald, 2013; Nozari & Novick, 2017). This additional cognitive demand may exert more pressure to efficiently use the limited memory resources in production than in comprehension, possibly resulting in a stronger effect of SRA. Future work is needed to investigate this possibility.

Implications for linguistic typology

Dependency length minimization as a functional universal has been argued to shape the syntactic structure of human language, due to the pressure to efficiently use the limited working memory resources. In the current study, we go beyond the general constraint of limited memory capacity, and further argue that memory resources should be strategically allocated to prioritize novel and unexpected information, a memory efficiency principle that naturally arises from two assumptions about working memory. Our results indicate that this strategic resource allocation indeed serves as a functional constraint to shape syntactic structures, in the sense that the pressure to minimize dependency length can actually be relaxed when the antecedent of a syntactic dependency is of higher surprisal. Our finding further substantiates the functionalist view as a promising approach to provide explanatory accounts for linguistic universals (Gibson et al., 2019). It also highlights the importance of having increasingly sophisticated characterization of functional constraints, in order to see how far we can go with this functionalist view on the structure of human language.

Despite this goal of having SRA as a universal efficiency principle to explain language structure, an important question is: how universal is SRA cross-linguistically, and how does it interact with other grammatical constraints and language-specific phrase structures?

First of all, one consistent pattern we have observed is the asymmetry between subject and object relations. Specifically, the effect of SRA is generally less reliable for object relations than for subject relations. One possible explanation is that object relations are subject to stronger grammatical constraints, with greater pressure to position the head and its dependent closer to each other. In support of this possibility, Keenan and Comrie (1977) identify an Accessibility Hierarchy as a linguistic universal, where noun phrases in the subject position are more readily relativized into relative clauses than those in the object position. Such constraints may bind object noun phrases and verbs more tightly, reducing the influence of SRA. Moreover, in many languages, grammatical agreement is common in subject–verb dependencies but absent in object–verb dependencies. Therefore, establishing subject relations may require more active grammatical computation, resulting in increased memory demand in processing and a stronger need for more strategic and efficient use of working memory resources.

Another notable pattern in Study 1 is that, compared to head-initial languages, most of the head-final languages in our analysis do not exhibit a reliable antecedent surprisal effect on dependency length. We speculate that this may be related to the tendency for argument dropping in head-final languages with SOV word order. From the perspective of dependency locality, SOV word order is associated with longer dependency lengths compared to SVO, which should theoretically impose higher memory costs and make it a less efficient structure. Despite this inefficiency, typologically, SOV is a word order commonly attested (Hammarström, 2016). As an explanation for this paradox, some studies find that arguments in SOV languages are often dropped, reducing the overall dependency length in actual language use (Levshina, 2025; Ueno & Polinsky, 2009). It is possible that by allowing speakers to drop arguments, the efficiency of SOV structure may already be significantly improved, obviating the need for SRA as another efficiency strategy.

It is also worth mentioning a few other relevant language-specific factors. For Mandarin Chinese, although it predominantly follows SVO word order, the relative clause goes before the nouns, increasing the flexibility of object relations in terms of their dependency length. For Korean and Japanese, the subject noun phrase may be delayed when it is surprising and unexpected, shifting the word order from SOV to OSV, and therefore decreasing the dependency length of subject relations for surprising antecedents.

Surprisal effect as efficient coding: An adaptionist view

One way to interpret our finding is that the enhanced robustness of memory representation arises as a by-product of the effort involved in processing surprising information. However, this raises an even more fundamental question: why does the surprisal effect occur in the first place? In other words, why is there a widely observed positive relationship between surprisal and processing effort?

Here is one way to think about the basic surprisal effect from an information-theoretic perspective. More surprising linguistic units correspond to longer code length. For example, consider a low-surprisal word encoded by a sequence of 3 bits `110`, compared to a high-surprisal word encoded by 11 bits `11100001101`. An (over-)simplified mechanical interpretation of the basic surprisal effect, therefore, is that encoding longer sequence of code in memory requires more time and effort. This leads to the widely observed linear relationship between surprisal and behavioral measures, such as reading time (e.g., Hoover et al., 2023; Shain et al., 2024; Smith & Levy, 2013; Wilcox et al., 2023; Xu et al., 2023).

However, choosing a code length based on predictability is not the only possible strategy to encode a linguistic unit. An alternative approach is to assign each unit a code of equal length regardless of statistical regularities, a scheme analogous to the uniform distribution of resource allocation discussed in Introduction. For example, the ASCII (American Standard Code for Information Interchange) system uses such a coding strategy, where every character is represented as a 7-bit sequence. Under this hypothetical scenario, processing effort would be insensitive to statistical regularities and uniformly distributed across all linguistic units.

As demonstrated in previous work on information theory, a coding scheme that treats every input as equally likely is inherently inefficient. When the statistical structure of the input is known, a coding scheme that assigns shorter code lengths to more predictable inputs

can reduce the average code length, thus increasing efficiency (Shannon, 1948). The observed linear relationship between surprisal and processing effort suggests that the brain may adopt such an efficient coding strategy. Specifically, the brain appears to assign code lengths to input units in proportion to their likelihood of occurrence based on long-term statistical regularities. This strategy is supported by the evidence that the brain is very good at inferring and approximating the statistical structure of the external environment (Saffran, Aslin, & Newport, 1996), and there is good reason to believe that the brain uses these inferred statistical structures to encode information more efficiently.

The connection between the surprisal effect and efficient coding suggests that the basic surprisal effect itself may be construed in evolutionary terms. Under certain efficiency pressures, the cognitive system may have evolved to adopt a memory encoding strategy that optimizes an objective function within the constraint of limited resources. The exact nature of this objective function remains an open question: it could be the minimization of distortion cost, or it could involve something else. The key point, however, is that the relationship between surprisal and strategic resource allocation can be understood at different timescales. On the one hand, at the level of processing individual sentences, strategic resource allocation may arise as a by-product of the effort required to process surprising information. On the other hand, over a longer timescale, the tendency to invest more effort into encoding less predictable information may reflect an evolved strategy that is adapted to an efficiency problem.

Limitations

One major limitation of our analysis is that the results heavily depend on the quality of surprisal measures generated from LLMs. In the current study, we conducted our analysis using the surprisal from two language models, namely GPT-3 and mGPT. As presented above, the two models do yield consistent pattern, alleviating the concern that our main result may be the artifact of any model-specific behavior. However, it still does not entirely rule out the issue with the accuracy of LLM surprisals, especially for low-resource languages that are under-represented in the training data of the models we used. This limitation may compromise the extensibility of our analysis to understudied languages, which are of particular interest from a typological perspective, and is particularly relevant for the unreliable effect we observed for some non-Indo-European languages.

In the end, even though contemporary LLMs can provide state-of-the-art probabilistic measures for linguistic data, it remains questionable to what extent it reflects the predictive processing in humans. Despite the correlation between the model-generated surprisals and the behavioral or neural responses in humans, many studies actually find that there still remain some critical patterns in human empirical data that cannot be fully accounted for solely by surprisals (Huang et al., 2024; van Schijndel & Linzen, 2021). Moreover, compared to humans, modern LLMs are far less constrained in terms of their memory capacity. This makes LLMs less likely to resemble the memory architecture in humans, or to capture the memory processes stemmed from the efficiency pressure exerted by the limited memory capacity (Oh & Schuler, 2023; Timkey & Linzen, 2023). However, this is not *necessarily* a limitation for the current study, since there are cases where a model with superhuman memory can provide probabilistic measures that more accurately reflect the statistical properties in the linguistic data without being confounded by the memory interference in the language model itself.

Conclusion

The current study proposes Strategic Resource Allocation (SRA) as an efficiency principle for memory encoding in sentence processing, which holds that working memory resources are strategically and dynamically allocated to prioritize novel and unexpected information. Theoretically, we argue that SRA is an efficient solution to the computational problem faced by working memory, that is, to maximize the retrieval accuracy of past inputs under the constraint of limited memory resources. Empirically, this principle predicts that the memory representation of more surprising linguistic units is more robust against interference and decay. We examined this prediction through naturalistic corpus data in the context of dependency locality from both the comprehension and the production side. In production, through the analysis of UD corpora in 10 languages, we indeed found that more surprising antecedents can tolerate longer dependency length, but the effect mostly exists within Indo-European and head-initial languages. This cross-linguistic variability therefore calls for a closer look into how SRA as a domain-general memory efficiency principle interacts with the language-specific phrase structure. In comprehension, through two English reading time corpora, we observed a similar reduced locality effect on retrieval difficulty for more surprising antecedents. Moreover, we found that the effect is more reliable for dependencies of subject relations than object relations. Taken together, there is converging evidence from naturalistic corpus data supporting that unpredictable antecedents are encoded with enhanced representation to be more resistant against memory decay and interference, a pattern that is predicted by our SRA.

CRediT authorship contribution statement

Weijie Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Richard Futrell:** Writing – review & editing, Validation, Supervision, Investigation, Funding acquisition, Data curation, 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. Mathematical derivation of strategic resource allocation

In this section, we demonstrate the mathematical derivation of strategic resource allocation. First, we will characterize the memory retrieval process as Bayesian inference for the encoded linguistic input (e.g., words). For the purpose of this derivation, we will assume that the linguistic prediction and the underlying memory representation follow Gaussian distributions; these may be interpreted as distributions over values of features. Needless to say, this is a highly simplified view of mental lexicon, and is not necessarily the reality for memory encoding, especially given that word inputs are discrete units rather than continuous variables. However, Gaussian distribution has some desirable mathematical properties with analytical solutions to help us validate the intuition behind our proposal.

A.1. Memory retrieval via Bayesian inference

Suppose one is trying to encode an input word w in noisy memory. We model noise by assuming that the input representation w is corrupted by Gaussian noise with an adjustable precision τ_w, yielding a noisy memory representation r:

$$r \sim w + \mathcal{N}\left(0, \tau_w^{-1}\right). \tag{A.1}$$

Retrieval from memory is then performed by forming a reconstructed representation $\hat{w} = \mathbb{E}_{p(w|r)}[w]$ as the posterior mean on input representations w given noisy memory representations r and a prior distribution on inputs p_0, with posterior distribution

$$p(\hat{w} \mid r) \propto p_M(r \mid w) p_0(w). \tag{A.2}$$

Now for ease of analysis, we set the prior distribution on input representations w to be a Gaussian distribution parameterized with mean w_0 and the precision τ_0:

$$p_0(w) = \mathcal{N}\left(w \mid w_0, \tau_0^{-1}\right). \tag{A.3}$$

A useful property of Gaussian distribution is that it forms a conjugate prior, so the posterior distribution on inputs words w given memory representations r is also Gaussian distribution. We assume that, in memory retrieval, the decoder[18]

$$\hat{w} \mid r \sim \mathcal{N}\left(\mu_{\text{post}}, \tau_{\text{post}}^{-1}\right), \tag{A.4}$$

where the posterior mean and precision are

$$\mu_{\text{post}} = \left(1 - \alpha_w\right) w_0 + \alpha_w r, \;\; \tau_{\text{post}} = \tau_0 + \tau_w, \tag{A.5}$$

where $\alpha_w = \frac{\tau_w}{\tau_0 + \tau_w}$. Then marginalizing out the memory representations, the distribution on reconstructed words is

$$\hat{w} \mid w \sim \mathcal{N}\left(\alpha_w w + \left(1 - \alpha_w\right) w_0, \frac{\alpha_w}{\tau_0 + \tau_w}\right). \tag{A.6}$$

We see that the retrieved word $\hat{w}$ is pulled towards the prior mode w_0 with a weight that depends on the encoding precision τ_w. As the encoding precision τ_w increases, this attraction to the prior is reduced.

Expected retrieval error under the memory model. We define the expected retrieval error $\varepsilon(w)$ for input w as the mean squared error between input w and reconstruction $\hat{w}$:

$$\varepsilon(w) = \mathbb{E}\left[(\hat{w} - w)^2 \mid w\right]. \tag{A.7}$$

We can express this mean squared error in terms of the bias–variance decomposition as

$$\varepsilon(w) = \text{Var}[\hat{w} \mid w] + \left(\mathbb{E}\left[\hat{w} \mid w\right] - w\right)^2. \tag{A.8}$$

Dropping in the mean and variance from Eq. (A.6), we can express the bias and variance as

$$\text{Var}\left[\hat{w} \mid w\right] = \frac{\alpha_w}{\tau_0 + \tau_w} \tag{A.9}$$

$$\mathbb{E}\left[\hat{w} \mid w\right] - w = \alpha_w w + \left(1 - \alpha_w\right) w_0 - w \tag{A.10}$$

$$= \left(1 - \alpha_w\right)\left(w_0 - w\right). \tag{A.11}$$

This gives us a convenient expression for the expected retrieval error,

$$\varepsilon(w) = \frac{\tau_w}{\left(\tau_w + \tau_0\right)^2} + \left(\frac{\tau_0}{\tau_0 + \tau_w}\left(w_0 - w\right)\right)^2 \tag{A.12}$$

$$= \frac{\tau_w + \tau_0^2\left(w - w_0\right)^2}{\left(\tau_0 + \tau_w\right)^2}. \tag{A.13}$$

Furthermore, we will wish to express the expected retrieval error in terms of the surprisal $h_w = -\ln p_0(w)$ of input w and the encoding precision τ_w. From the assumed Gaussian form of the prior over input words, we have

$$p_0(w) = \sqrt{\frac{\tau_0}{2\pi}} \exp\left(-\frac{\tau_0 (w - w_0)^2}{2}\right) \tag{A.14}$$

$$h_w = \frac{\tau_0 (w - w_0)^2}{2} - \ln\sqrt{\frac{\tau_0}{2\pi}}. \tag{A.15}$$

Extracting $\tau_0 (w - w_0)^2$ to the left-hand side yields:

$$\tau_0 (w - w_0)^2 = 2h_w + \ln\frac{\tau_0}{2\pi}. \tag{A.16}$$

Substituting this back into the expression for expected retrieval error, we get

$$\varepsilon(h_w, \tau_w) = \frac{\tau_w + 2\tau_0 h_w + \tau_0 \ln\frac{\tau_0}{2\pi}}{\left(\tau_0 + \tau_w\right)^2}. \tag{A.17}$$

Below, we will consider how to choose encoding precisions in order to minimize the expected retrieval error under constraints.

A.2. minimizing expected error under memory constraint

In this section, we aim to show that strategic resource allocation arises as a solution to the problem of minimizing the expected retrieval error on average:

- **Strategic Resource Allocation**
  Given two linguistic inputs, the minimization of their total expected error bounded by certain memory constraint requires that the more surprising input be encoded with higher precision.

Consider two input words w_1 and w_2 to be encoded and retrieved, whose surprisals are $h_{w_1} = -\ln p_0(w_1)$ and $h_{w_2} = -\ln p_0(w_2)$ respectively. We assume that the encoding precision τ_w for each word is proportional to the memory resources allocated, and we assume that there is a constraint on total memory resources c allocated for both words, which is to be distributed between w_1 and w_2. That is, we posit a constraint on the sum of encoding precisions $\tau_{w_1} + \tau_{w_2} = c$.

We will show that this optimization problem bounded by memory constraint leads to strategic resource allocation in memory encoding, as stated in the Proposition below:

Proposition. To minimize the total expected retrieval error for two linguistic inputs $\varepsilon_{w_1} + \varepsilon_{w_2}$ subject to a resource constraint $\tau_{w_1} + \tau_{w_2} = c$, the input that is more surprising under the prior distribution must be encoded with higher precision. Specifically, if

$$h_{w_1} > h_{w_2}, \tag{A.18}$$

then the optimal encoding satisfies

$$\tau_{w_1} > \tau_{w_2}. \tag{A.19}$$

Proof. First, in order to minimize expected error ε, we take the derivative of ε with respect to encoding precision τ_w for each input[19]:

$$\frac{\partial \epsilon(\tau_w, h_w)}{\partial \tau_w} = \frac{\tau_0 - \tau_w - 2\tau_0^2\left(w - w_0\right)^2}{\left(\tau_0 + \tau_w\right)^3}, \tag{A.20}$$

which reveals three possible situations with respect to the monotonicity of ε for both inputs:

1. The expected error monotonically decreases with increasing τ_w within its meaningful domain (i.e., $\tau_w > 0$) for both inputs w_1 and w_2. As shown below, this is the situation for most cases where the input w is not too close to the prior prediction w_0 and the prior precision τ_0 is not too unreliable.
2. The expected error is a non-monotonic function of τ_w for both inputs w_1 and w_2.
3. The expected error monotonically decreases with τ_w within in meaningful domain for one input but is non-monotonic for the other.

[18] This assumes that the decoder has access to the encoding precision τ_w for the input.

[19] For simplicity, we maintained $(w - w_0)^2$ for now in Eq. (A.20) instead of having it transformed to the form that contains the surprisal h_w of the input.

In this proof, we will show that the proposition above holds in all these three situations.

Situation 1. In this first situation, the expected error ϵ monotonically decreases for both inputs, which means that

$$\frac{\partial \epsilon(\tau_w, h_w)}{\partial \tau_w} \leq 0. \tag{A.21}$$

Thus, Situation 1 holds when

$$\tau_0 - \tau_w - 2\tau_0^2 \left(w - w_0\right)^2 \leq 0 \tag{A.22}$$

$$\tau_w \geq \tau_0 - 2\tau_0^2 \left(w - w_0\right)^2. \tag{A.23}$$

Since $\tau_w > 0$, the inequality in Eq. (A.23) always hold within the meaningful domain of τ_w if

$$\tau_0 - 2\tau_0^2 \left(w - w_0\right)^2 \leq 0 \tag{A.24}$$

$$\tau_0 \geq \frac{1}{2\left(w - w_0\right)^2}. \tag{A.25}$$

Intuitively, this means that the expected error ϵ monotonically decreases if the input w is not too close to the prior prediction w_0 and the prior precision τ_0 is not too unreliable.

Now that ϵ monotonically decreases within the meaningful domain of τ_w for both inputs w_1 and w_2 as defined in this first situation, using Eq. (A.17), we evaluate the derivative with respect to its relationship with input surprisal

$$\begin{aligned} f(\tau_w, h_w) &= \frac{\partial \epsilon(\tau_w, h_w)}{\partial \tau_w} \\ &= \frac{\tau_0 - \tau_w - 2\tau_0 \left(2h(w) + \ln\left(\frac{\tau_0}{2\pi}\right)\right)}{(\tau_0 + \tau_w)^3} \\ &< 0 \text{ (by the definition of Situation 1),} \end{aligned} \tag{A.26}$$

which shows that $f(\tau_w, h_w)$ monotonically decreases in h_w.

When $h_{w_1} = h_{w_2}$, $f(\tau_{w_1}, h_{w_1})$ and $f(\tau_{w_2}, h_{w_2})$ have the same quantitative form. As a result, the memory resources will be uniformly distributed across w_1 and w_2 with no impetus to redistribute more resources to any one of them:

$$\tau_{w_1} = \tau_{w_2} = \frac{c}{2}. \tag{A.27}$$

However, when $h_{w_1} > h_{w_2}$, since $f(\tau_w, h_w)$ monotonically decreases in h_w, we have $f(\tau_{w_1}, h_{w_1}) < f(\tau_{w_2}, h_{w_2})$ at any given value of τ_w. As a result, compared to the uniform distribution in Eq. (A.27), there is reason to redistribute more resources to the high surprisal w_1, since the decrease of error on w_1 will be higher than the increase of error on w_2, yielding a lower total error across the two inputs. Therefore, in Situation 1, if $h_{w_1} > h_{w_2}$, the optimal encoding strategy should satisfy $\tau_{w_1} > \tau_{w_2}$ (see Fig. A.1).

Situation 2. As shown above, in order for ϵ to be a non-monotonic function of τ_w:

$$\tau_0 < \frac{1}{2\left(w - w_0\right)^2}, \tag{A.28}$$

which corresponds to borderline cases where the input w is too close to the prior prediction w_0 or the prior precision τ_0 is too unreliable.

In this second situation, with increasing τ_w, ϵ first increases and then decreases, as illustrated in Fig. A.2. And the relationship between high surprisal and low surprisal inputs has three phases. We will show that the proposition still holds for all the three phases in Situation 2, such that the optimal encoding satisfies $\tau_{w_1} > \tau_{w_2}$ if $h_{w_1} > h_{w_2}$.

In **Phase III**, the situation is basically the same as the Situation 1 discussed above, where ϵ decreases with increasing τ_w for both inputs. Therefore, according to the proof in Situation 1, the optimal encoding satisfies $\tau_{w_1} > \tau_{w_2}$ if $h_{w_1} > h_{w_2}$.

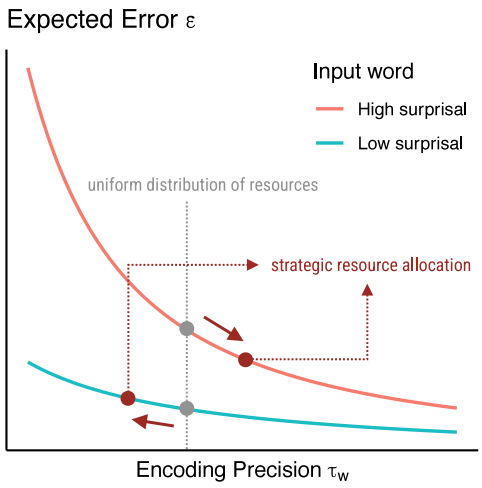


Fig. A.1. Situation 1 expected retrieval error ϵ for two input words as a function of encoding precision τ_w and their surprisal.

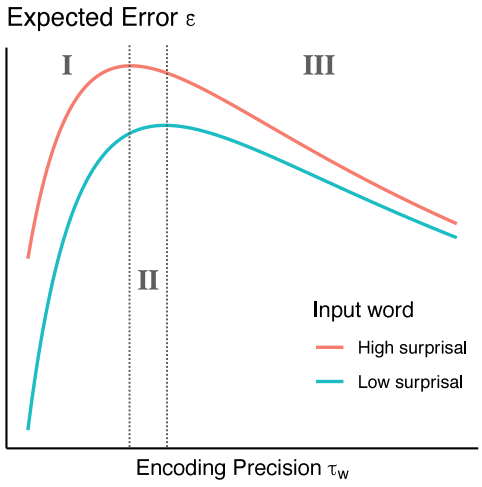


Fig. A.2. Situation 2 expected retrieval error ϵ for two input words as a function of encoding precision τ_w and their surprisal.

In **Phase II**, ϵ decreases in τ_w for one input and increases for the other. For each input, the turning point where the monotonicity is flipped is at

$$\tau_w = \tau_0 - 2\tau_0 \left(2h(w) + \ln\left(\frac{\tau_0}{2\pi}\right)\right), \tag{A.29}$$

where the derivative of ϵ in Eq. (A.26) is 0. Importantly, if $h_{w_1} > h_{w_2}$, then the turning points $\tau_{w_1} < \tau_{w_2}$. Therefore, the turning point for the high surprisal input w_1 is to the left of the one for the low surprisal input w_2. That means, in Phase II, it can only be the case that the expected error ϵ decreases in τ_w for the high surprisal w_1, but increases for the low surprisal w_2. As a result, in order to minimize ϵ, more memory resources should be allocated to encode w_1 than w_2, leading to $\tau_{w_1} > \tau_{w_2}$.

In **Phase I**, ϵ increases in τ_w for both inputs w_1 and w_2. Recall that the derivative of ϵ (repeated below in Eq. (A.30)) decreases as the input surprisal h_w increases.

$$\begin{aligned} f(\tau_w, h_w) &= \frac{\partial \epsilon(\tau_w, h_w)}{\partial \tau_w} \\ &= \frac{\tau_0 - \tau_w}{(\tau_0 + \tau_w)^3} - \frac{2\tau_0 \left(2h(w) + \ln\left(\frac{\tau_0}{2\pi}\right)\right)}{(\tau_0 + \tau_w)^3} \\ &> 0. \end{aligned} \tag{A.30}$$

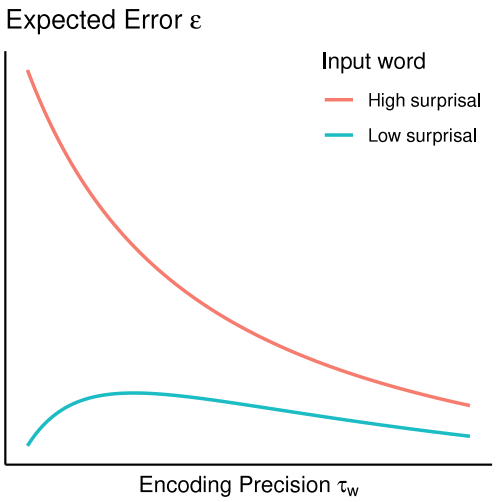


Fig. A.3. Situation 3 expected retrieval error ϵ for two input words as a function of encoding precision τ_w and their surprisal.

Therefore, in Phase I, the expected error ϵ increases more slowly for high surprisal input. As a result, if $h_{w_1} > h_{w_2}$, then for a fixed amount of memory resources, more resources allocated to w_1 yields lower smaller increase in ϵ, leading to $\tau_{w_1} > \tau_{w_2}$.

Situation 3. In this third situation, the expected error ϵ monotonically decreases in τ_w for one input, but is a non-monotonic function for the other.

Recall that whether ϵ is monotonic depends on the inequality in Eq. (A.28). That is, in order for ϵ to be non-monotonic, Eq. (A.28) must hold. Moreover, as discussed in Situation 2 above, if $h_{w_1} > h_{w_2}$, the turning point of monotonicity for the high surprisal input w_1 is to the left of the one for the low surprisal input w_2. As a result, in this Situation 3, it must be the case that it is the high surprisal input w_1 that monotonically decreases in τ_w whereas the low surprisal input w_2 first increases and then decreases, as illustrated in Fig. A.3.

Apparently, Situation 3 is basically equivalent to the Phase II and Phase III in Situation 2. Therefore, as proved above, if $h_{w_1} > h_{w_2}$, the optimal encoding should satisfy $\tau_{w_1} > \tau_{w_2}$. □

Remark. To sum up, if the surprisal of two input words $h_{w_1} > h_{w_2}$, given fixed amount of memory resources such that the encoding precisions for two inputs is constrained by $\tau_{w_1} + \tau_{w_2} = c$, the optimal encoding with strategic resource allocation should satisfy $\tau_{w_1} > \tau_{w_2}$ in order to achieve minimal total expected error ε. We outlined three possible situations of how ϵ may change with increasing encoding precision τ_w, and we proved that the strategic resource allocation should hold in all three situations. It is worth noting that, in most cases, ϵ monotonically decreases with increasing encoding precision τ_w, as in Situation 1. However, when the prior precision is too unreliable or when the input word is too close to the prior predicted word, there will be borderline cases where ϵ first increases with increasing τ_w before it starts to decrease, as in Situation 2 and 3.

Appendix B. Statistical models

B.1. Study 1

In Study 1, for each language, we ran regression models on dependency length L. As mentioned in the main article, the regression models were run separately for L measured as intervening word counts and as intervening surprisal, as shown below in (1) and (2). For the analysis on full dataset, we ran linear mixed-effect model with random intercept per dependency type. For analysis on subject relations and object relations, we ran the standard linear regression without specifying random effects. Compare to the orthographic L, the analysis with information-theoretic L includes BASELINE SURPRISAL as an additional control variable.

(1) Regression formulas for orthographic L_{O}

- *Full dataset*
  $L \sim 1$ + Sentence Position + Antecedent Position + Sentence Length + Antecedent Frequency + Antecedent Surprisal + (1 | Dependency Type)
- *Subject/object relations*
  $L \sim 1$ + Sentence Position + Antecedent Position + Sentence Length + Antecedent Frequency + Antecedent Surprisal

(2) Regression formulas for information-theoretic L_{I}

- *Full dataset*
  $L \sim 1$ + Sentence Position + Antecedent Position + Sentence Length + Baseline Surprisal + Antecedent Frequency + Antecedent Surprisal + (1 | Dependency Type)
- *Subject/object relations*
  $L \sim 1$ + Sentence Position + Antecedent Position + Sentence Length + Baseline Surprisal + Antecedent Frequency + Antecedent Surprisal

B.2. Study 2a

In Study 2a, we ran linear mixed-effect models on log-transformed reading times for the critical region at the retrieval site (i.e., the right codependent for each syntactic dependency) and its spillover region.

(3) Regression formulas for the critical region

a. Fixed effects
   logRT ~ 1 + sent.pos + antec.pos + sent.len + word.len + antec.freq + surp + surp.prev1 + surp.prev2 + freq + freq.prev1 + freq.prev2 + L * antec.surp
b. Random effects
   - Orthographic L_{O}
     - *Full dataset*: (1 | dep.type) + (1 | part)
     - *Subject relations*: (L + antec.surp | part)
     - *Object relations*: (L + antec.surp | part)
   - Info-theoretic L_{I}
     - *Full dataset*: (1 | dep.type) + (1 | part)
     - *Subject relations*: (antec.surp | part)
     - *Object relations*: (L + antec.surp | part)

(4) Regression formulas for the spillover region

a. Fixed effects
   logRT ~ 1 + sent.pos + antec.pos + sent.len + word.len + antec.freq + surp + surp.prev1 + surp.prev2 + freq + freq.prev1 + freq.prev2 + L * antec.surp
b. Random effects
   - Orthographic L_{O}
     - *Full dataset*: (1 | dep.type) + (1 | part)
     - *Subject relations*: (L * antec.surp | part)
     - *Object relations*: (antec.surp | part)
   - Info-theoretic L_{I}
     - *Full dataset*: (1 | dep.type) + (1 | part)
     - *Subject relations*: (antec.surp | part)
     - *Object relations*: (antec.surp | part)

B.3. Study 2b

In Study 2b, we ran linear mixed-effect models on first-pass durations and total reading times at the retrieval site, as shown below in (5) and (6).

(5) Regression formulas for first-pass durations (same maximal converging random structure for both measures of L)

a. Fixed effects
logRT ~ 1 + sent.pos + antec.pos + sent.len + word.len + antec.freq + surp + surp.prev1 + surp.prev2 + freq + freq.prev1 + freq.prev2 + L * antec.surp
b. Random effects

- *Full dataset*: (L * antec.surp | dep.type) + (L * antec.surp | part)
- *Subject relations*: (L * antec.surp | part)
- *Object relations*: (L * antec.surp | part)

(6) Regression formulas for total reading times (same maximal converging random structure for both measures of L in the analysis of full dataset and subject relations)

a. Fixed effects
logRT ~ 1 + sent.pos + antec.pos + sent.len + word.len + antec.freq + surp + surp.prev1 + surp.prev2 + freq + freq.prev1 + freq.prev2 + L * antec.surp
b. Random effects

- *Full dataset*: (L * antec.surp | dep.type) + (L * antec.surp | part)
- *Subject relations*: (L * antec.surp | part)
- *Object relations*:
  - Orthographic L_O: (L * antec.surp | part)
  - Info-theoretic L_I: (L + antec.surp | part)

Data availability

The analysis code is available at: https://osf.io/yf4ca/.

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.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

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.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(5), 1178.

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

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., Catalao, R. F., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7.

Bays, P. M., Schneegans, S., Ma, W. J., & Brady, T. F. (2024). Representation and computation in visual working memory. *Nature Human Behaviour*, 1–19.

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.

Bosco, C., Montemagni, S., & Simi, M. (2013). Converting Italian treebanks: Towards an Italian stanford dependency treebank. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse* (pp. 61–69).

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., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity: More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences*, *113*(27), 7459–7464.

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.

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.

Ferrer-i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*(5), Article 056135.

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.

Chun, J., Han, N.-R., Hwang, J. D., & Choi, J. D. (2018). Building universal dependency treebanks in Korean. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Clark, T. H., Meister, C., Pimentel, T., Hahn, M., Cotterell, R., Futrell, R., et al. (2023). A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, *11*, 1048–1065.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In *Advances in psychology*: *vol. 9*, (pp. 287–299). Elsevier.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

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. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Droganova, K., Lyashevskaya, O., & Zeman, D. (2018). Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (TLT 2018)*: *vol. 155*, (pp. 53–66).

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

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, *37*(2), 378–394.

Ferreira, V. S. (2003). 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, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology*, *70*, 29–51.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 203–218.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

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. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the first workshop on quantitative syntax (quasy, syntaxFest 2019)* (pp. 2–15).

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., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., et al. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*, 63–77.

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the fifth international conference on dependency linguistics (depling, syntaxFest 2019)* (pp. 3–13).

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*(7), 615–621.

Gahl, S., & Baayen, R. H. (2024). Time and thyme again: Connecting english spoken word duration to models of the mental lexicon. *Language*, *100*(4), 623–670.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons, Behavior, Data Analysis, and Theory*, *2*(3), 1–10.

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.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., et al. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59*(1), 23–59.

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).

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing.. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(6), 1411.

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

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.

Hahn, M., & Xu, Y. (2022). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences*, *119*(24), Article e2122604119.

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*.

Hammarström, H. (2016). Linguistic diversity and language evolution. *Journal of Language Evolution*, *1*(1), 19–29.

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.

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*(6), 304–313.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry*, *21*(2), 223–261.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. (No. 73), Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hedayati, S., O'Donnell, R. E., & Wyble, B. (2022). A model of working memory for latent representations. *Nature Human Behaviour*, *6*(5), 709–719.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews. Neuroscience*, *13*(2), 135–145.

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, K.-J., & Dillon, B. (2023). An infrequent, large cost underlies garden path effects: An RT distribution approach. In *The 36th annual conference on human sentence processing* (pp. 9–11).

Icard, T. (2023). Resource rationality. In *Book manuscript: vol. 434*.

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., & Buz, E. (2017). Signal reduction and linguistic encoding. In *The Handbook of Psycholinguistics* (pp. 38–81). Wiley Online Library.

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.

Jakob, A. M., & Gershman, S. J. (2023). Rate-distortion theory of neural coding and its implications for working memory. *Elife*, *12*, Article e79450.

Johannsen, A., Alonso, H. M., & Plank, B. (2015). Universal dependencies for danish. In *International workshop on treebanks and linguistic theories* (p. 157).

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*(1), 122.

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

Karimi, H., & Ferreira, F. (2016). Informativity renders a referent more accessible: Evidence from eyetracking. *Psychonomic Bulletin & Review*, *23*, 507–525.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168.

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

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

Koranda, M. J., Bulgarelli, F., Weiss, D. J., & MacDonald, M. C. (2020). Is language production planning emergent from action planning? A preliminary investigation. *Frontiers in Psychology*, *11*, 1193.

Kowialiewski, B., Lemaire, B., & Portrat, S. (2022). Between-item similarity frees up working memory resources through compression: A domain-general property. *Journal of Experimental Psychology: General*, *151*(11), 2641.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).

Levshina, N. (2025). The paradox of SOV: A case for token-based typology. http://dx.doi.org/10.31234/osf.io/wfbpv, PsyArXiv.

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., & 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.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382–1411.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.
Liu, Z. (2020). Mixed evidence for crosslinguistic dependency length minimization. *STUF-Language Typology and Universals*, *73*(4), 605–633.
Liu, Z. (2021). The crosslinguistic relationship between ordering flexibility and dependency length minimization: A data-driven approach. *Society for Computation in Linguistics*, *4*(1).
Liu, Z., & Wulff, S. (2023). The development of dependency length minimization in early child language: A case study of the dative alternation. In *Proceedings of the seventh international conference on dependency linguistics (depling, GURT/syntaxFest 2023)* (pp. 1–8).
Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*(3), 550–557.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.
Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.
Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.
MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.
MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on just and carpenter (1992) and waters and caplan (1996). *Psychological Review*, *109*(1).
Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.
Marr, D. (1982). *Vision: A computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co. Inc..
Marşan, B., Akkurt, S. F., Şen, M., Gürbüz, M., Güngör, O., Özateş, Ş. B., et al. (2022). Enhancements to the BOUN treebank reflecting the agglutinative nature of turkish. arXiv preprint arXiv:2207.11782.
McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 92–97).
Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 963–980).
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.
Miller, G. A., & Isard, S. (1964). Free recall of self-embedded english sentences. *Information and Control*, *7*(3), 292–303.
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.
Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., et al. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4034–4043).
Nozari, N., & Novick, J. (2017). Monitoring and control in language production. *Current Directions in Psychological Science*, *26*(5), 403–410.
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.
Opedal, A., Chodroff, E., Cotterell, R., & Wilcox, E. (2024). On the role of context in reading time prediction. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 3042–3058).
Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review*, *105*(4), 761.
Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*, 369–378.
Pearlmutter, N. J., & Gibson, E. (2001). Recency in verb phrase attachment.. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *27*(2), 574.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.
Pierrehumbert, J. B. (2008). Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the emergence of linguistic structure* (pp. 137–158). John Benjamins Publishing Company.
Pimentel, T., Meister, C., Wilcox, E., Mahowald, K., & Cotterell, R. (2023). Revisiting the optimality of word lengths. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 2240–2255). Association for Computational Linguistics.
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.
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.
Rathi, N., Hahn, M., & Futrell, R. (2021). An information-theoretic characterization of morphological fusion. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10115–10120). Association for Computational Linguistics.
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.
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.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.
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. *Bell System Technical Journal*, *27*(3), 379–423.
Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., & Shavrina, T. (2024). mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, *12*, 58–79.
Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99–118.
Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.
Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory.. *Psychological Review*, *119*(4), 807.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.
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.
Speer, R. (2022). rspeer/wordfreq: v3.0. http://dx.doi.org/10.5281/zenodo.7199437.
Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., et al. (2016). Universal dependencies for Japanese. In *Proceedings of the tenth international conference on language resources and evaluation* (pp. 1651–1658).
Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of 6th international conference on language resources and evaluation: vol. 2008*, (pp. 96–101).
Temperley, D., & Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, *4*, 67–80.
Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: bridging the gap between computational and empirical approaches to reference*.
Timkey, W., & Linzen, T. (2023). A language model with limited memory capacity captures interference in human sentence processing. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 8705–8720).
Traugott, E. C., & Trousdale, G. (2013). *Constructionalization and constructional changes*: *vol. 6*, OUP Oxford.
Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90.
Troyer, M., Hofmeister, P., & Kutas, M. (2016). Elaboration over a discourse facilitates retrieval in sentence processing. *Frontiers in Psychology*, *7*, 374.
Ueno, M., & Polinsky, M. (2009). Does headedness affect processing? A new look at the VO–OV contrast1. *Journal of Linguistics*, *45*(3), 675–710.
van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, Article e34963.
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.
Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in pavlovian conditioning: Application of a theory. *Inhibition and Learning*, 301–336.

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.

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. (2024). A hierarchical Bayesian model for syntactic priming. In *Proceedings of the annual meeting of the cognitive science society*: *vol. 46*.

Xu, W., & Futrell, R. (2025). Informativity enhances memory robustness against interference in sentence comprehension. *Journal of Memory and Language*, *142*, Article 104603.

Xu, W., & Xiang, M. (2021). Is there a predictability hierarchy in reference resolution? In *Proceedings of the annual meeting of the cognitive science society*: *vol. 43*, (43).

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466.

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.