Lossy Context Surprisal Predicts Task-Dependent Patterns in Relative Clause Processing

Kate McCurdy and **Michael Hahn**
Universität des Saarlandes

Abstract

English relative clauses are a critical test case for theories of syntactic processing. Expectation- and memory-based accounts make opposing predictions, and behavioral experiments have found mixed results. We present a technical extension of Lossy Context Surprisal (LCS) and use it to model relative clause processing in three behavioral experiments. LCS predicts key results at distinct retention rates, showing that task-dependent memory demands can account for discrepant behavioral patterns in the literature.

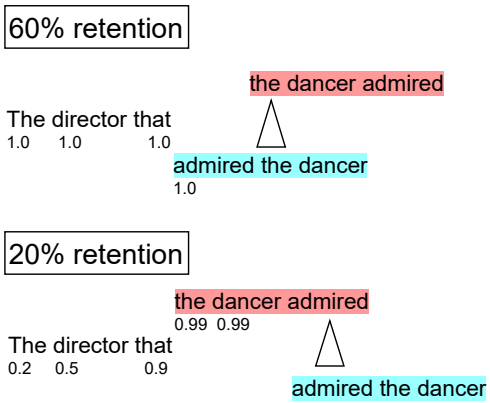

Figure 1: Illustration of lossy context surprisal (LCS) with retention probabilities of individual words. At high retention rates (top), LCS predicts an expectation-based processing slowdown at “the" for object relative clauses (red). At low retention rates (bottom), LCS predicts a memory-based processing slowdown at the verb.

1 Introduction

A fundamental goal of computational psycholinguistics is to predict and explain syntactic processing difficulty as manifested in reading times. Surprisal from modern language models is a strong predictor of reading times on naturalistic text: words take longer to read when they are less predictable (e.g. Wilcox et al., 2023). This finding aligns with expectation-based theories of syntactic processing (Hale, 2001; Levy, 2008). However, surprisal fails to account for certain effects from the psycholinguistic literature — particularly *locality effects*, in which longer syntactic dependencies lead to increased processing effort (e.g. Grodner and Gibson, 2005; Bartek et al., 2011). Under surprisal theory, this is unexpected: additional intervening context should generally make prediction easier.

Locality effects are naturally explained in terms of human memory limitations, which motivate memory-based theories of syntactic processing. One example is Dependency Locality Theory (Gibson, 1998; Gibson et al., 2000), which posits that the processing cost of integrating a syntactic dependency is proportional to dependency length. Similar locality predictions arise from cue-based retrieval theories (e.g. Lewis and Vasishth, 2005).

Recent research has offered a principled conceptual unification of expectation- and memory-based perspectives in terms of *Lossy-Context Surprisal* (LCS; Futrell et al., 2020). This theory holds that expectations are derived from imperfect memory representations of the context; hence, words are easy to process only when they are easy to predict from lossy context representations. Resource-Rational Lossy-Context Surprisal (RR-LCS) (Hahn et al., 2022) implements LCS for general input by constraining GPT-2 (Radford et al., 2019) with rationally optimized lossy context representations.

Here, we use LCS to model memory and expectation in the context of English relative clause processing (Figure 1) – long considered a key setting where memory- and expectation-based models make opposing predictions (e.g. Levy, 2008, 2013). Object relative clauses (ORCs), such as “The director that the dancer admired," are more difficult to process than subject relative clauses (SRCs), such as “The director that admired the dancer." Surprisal theory and DLT differ as to *when* this

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 36–45

difficulty arises in incremental processing. Under the expectation-based account of surprisal theory, comprehenders use their experience of English syntactic distributions to predict upcoming structures. Subject relatives are more frequent than object relatives in written corpora (Roland et al., 2007). Therefore, given the prefix "The director that," readers should expect a tensed verb such as "admired," and slow down on encountering the ORC determiner "the." Surprisal theory thus predicts that the processing difficulty for ORCs relative to SRCs will appear primarily on the determiner. By contrast, DLT posits that processing difficulty reflects the integration of long-distance dependencies. Under this account, the main slowdown in ORCs should instead appear at the verb "admired," as comprehenders integrate the dependency to the distant object "director."

Behavioral studies of relative clause processing have found discrepant results depending on the task. Experimental data from eye-tracking (Traxler et al., 2002; Staub, 2010) and self-paced reading (Grodner and Gibson, 2005; Roland et al., 2007; Frinsel and Christiansen, 2024) support the memory-based prediction of longer reading time on ORC verbs. Using the Maze task, however, Forster et al. (2009) find only the determiner slowdown predicted by surprisal theory. In a recent study, Vani et al. (2021) collect Maze data with stimuli from earlier eye-tracking experiments, and reproduce the determiner slowdown. The authors suggest that the later ORC verb slowdown found in eye-tracking studies may reflect spillover rather than memory effects.

In the current study, we investigate whether the task-dependent discrepancies observed in English relative clause processing can be modeled as a trade-off between memory and expectation. We manipulate how much of the preceding sentence context is remembered in the lossy context surprisal model, and evaluate Vani et al.'s stimuli at a range of retention rates. We additionally evaluate LCS predictions on the relative clause stimuli of Roland et al. (2021), who report both spillover and memory effects in their eye-tracking data.

Figure 1 illustrates our results. At a high retention rate (e.g. 60%), LCS predicts the expectation-based determiner slowdown on ORC test items, consistent with the observed RTs for the Maze filler data. At a low retention rate (e.g. 20%), however, LCS predicts the ORC verb slowdown found in eye-tracking studies such as Staub (2010). Furthermore, we find that low-retention LCS predictions also capture the memory effects found by Roland et al. after adjusting for spillover per their analysis. This finding suggests an alternative explanation for observed task discrepancies: eye-tracking while reading likely imposes lower memory demands than the Maze task, leading to a stronger influence of memory constraints on incremental processing.

This paper presents two key contributions:[1]

- We release and document a technical improvement to the RR-LCS model. Through extending the lossy context model to subword tokenization, the new model can now handle out-of-vocabulary inputs.

- We show that, through manipulating the retention rate, LCS predicts two distinct behavioral patterns of relative clause processing which have been reported in different tasks. This finding shows that task-dependent memory demands can explain apparently contradictory results in the literature.

2 Background

Measuring incremental processing Behavioral methods which track word-by-word reading time (RT) offer scientific insight into human language processing, as longer RTs reflect processing difficulty. Special eye-tracking (ET) equipment can collect RT data in a laboratory setting by monitoring participants' eye movements as they read (Rayner, 1998). This method most closely approximates natural reading, but ET data collection is resource-intensive and the resulting RTs can be noisy and challenging to interpret. One crucial source of noise comes from *spillover effects*: longer processing time for one word can "spill over" to following words. In such cases, systematically longer RTs on a specific word do not reflect difficulty processing that word, but instead the word or words preceding.

An alternative cost-effective source of RT data is self-paced reading (SPR), in which participants must press a button to reveal each word in sequence. Unfortunately, spillover effects are typically much larger in SPR compared to ET data. The Maze task (Forster et al., 2009) modifies SPR by introducing distractors: participants are shown two words at each step, and must select the word which correctly continues the sentence. This task is more cognitively demanding, and appears to reduce or

[1] See https://github.com/kmccurdy/LCS for model and analysis code.

eliminate spillover effects (Boyce and Levy, 2020; Boyce et al., 2020). Witzel et al. (2012) compare Maze and ET for three types of ambiguous sentences and find that Maze RTs capture most — but not all — patterns of incremental processing difficulty seen in ET RTs. In this paper, we consider the possibility that higher working memory demands in the Maze task account for key discrepancies between Maze and ET results.

Modeling memory and expectation Language models (LMs) are typically trained on a next-word prediction objective, which aligns them with the expectation-based account of Surprisal Theory. Modern large language models, however, have become worse predictors of human RT data due to their superhuman capacity for memorization (Oh and Schuler, 2023). This has motivated modeling approaches which combine LMs with memory constraints. Timkey and Linzen (2023) propose a model architecture with a single self-attention head, which reduces the capacity to retrieve earlier representations from context. Kuribayashi et al. (2022) find improved fits to RTs by simply truncating words from the preceding context. Here, we model memory constraints with Resource-Rational Lossy Context Surprisal (RR-LCS; Hahn et al., 2022), which learns to stochastically retain or delete specific words from the representation of the preceding context. Crucially, we can systematically vary the LCS retention rate to simulate different patterns of working memory engagement.

3 Computing Lossy Context Surprisal

3.1 Resource-Rational Lossy Context Surprisal

Standard surprisal theory assumes that processing difficulty of a word is proportional to its surprisal—that is, its negative log-probability in context:

$$-\log P(x_{T+1}|x_{1\ldots T}) \tag{1}$$

Lossy Context Surprisal (Futrell et al., 2020) modifies this by conditioning not on the exact context, but on a lossy memory representation:

$$-\log P(x_{T+1}|M_T) \tag{2}$$

where M is a lossy representation generated from $x_1 \ldots x_T$. To generate testable Lossy Context Surprisal predictions, we must specify (1) lossy representations M_T and (2) how these are generated from contexts $x_{1\ldots T-1}$. Such a specification is provided by Resource-Rational Surprisal (RR-LCS; Hahn et al., 2022). Following Futrell et al. (2020), RR-LCS specifies the lossy representations in terms of retaining or masking individual words. Formally, the model operates over contexts $x \in \Sigma^T$, where T is a maximum context size, set to 20 in Hahn et al. (2022). The model is specified by a family of retention probabilities (after Anderson and Milson, 1989; Anderson and Schooler, 1991) $p_{w,i} \in [0, 1]$ $(1 \leq i \leq T)$, where $p_{w,i}$ indicates the probability that word w at position i is available when predicting word T (Figure 1). Given a context $x_{1\ldots T}$, each word is independently kept or masked depending on these probabilities, yielding a lossy representation $M_T := y \in (\Sigma \cup \{\text{LOST}\})^T$.

The retention probabilities $p_{w,i}$ are chosen so as to minimize average lossy-context surprisal:

$$\min_{p_{w,i}} \mathbb{E}_{x_{1\ldots T+1}, y_{1\ldots T}}[-\log P(x_{T+1}|y_1 \ldots y_T)] \tag{3}$$

subject to a bound on the average number of retained words:

$$\mathbb{E}_{x,y}[\#\{i : y_i = \text{LOST}\}] \leq \delta T \tag{4}$$

where the expectations range over contexts $x_{1\ldots T}$ with associated next word x_{T+1} from a large corpus, and lossy versions y drawn via the retention probabilities $p_{w,i}$. Importantly, the *retention rate* $\delta \in [0, 1]$ is the model's single free parameter: it indicates how many words on average are retained. Given a budget specified by δ, the model thus learns to prioritize retaining those words that are usually more helpful for predicting future words. On a technical level, the constrained optimization (3–4) is implemented using Lagrangian duality; see Hahn et al. (2022, Supp. Mat. §1) for details. Empirically, the optimized retention probabilities strongly favor forgetting less recent words, especially high-frequency function words.

3.2 Implementation

In the parameterization of Hahn et al. (2022), given the embedding g_i of the i-the token and p_i of the i-th position, the retention probabilities receive a log-biaffine parameterization after Dozat and Manning (2017):

$$p_{w,i} = \sigma\left(F p_i + MLP_2(g_i) + p_i^T MLP_1(g_i)\right) \tag{5}$$

where MLP_i denotes ReLU MLPs with one hidden layer with d dimensions, and σ is the logistic sigmoid function. Both the positional and word embeddings can directly influence the probability

(first and second summands); there is also an option for multiplicative interaction between the two (third summand). The parameters of the two MLPs, the transform F, and the embeddings g_i, and p_i are trainable parameters, optimized for (3–4).

By Bayes' Rule, the predictive distribution $P(x_{T+1}|M_T)$ in (2) is proportional to:

$$\sum_{x_1 \ldots x_T \in \Sigma^T} P(x_{1\ldots T+1}) P(M_T|x_{1\ldots T}) \quad (6)$$

where the sum ranges over hypothetical contexts $x_{1\ldots T}$, weighted by their probability of giving rise to the imperfect representation M_T. The term $P(M_T|x_{1\ldots T})$ can be computed in terms of $p_{w,i}$. The other term, $P(x_{1\ldots T+1})$, describes the expectations in the absence of any memory limitations; Hahn et al. (2022) estimate it using GPT-2 Medium (Radford et al., 2019). Plugging these components into (6), lossy-context surprisal (2) is then estimated using importance sampling. Importantly, in the limit where no memory limitations are present ($\delta = 1$), the predictions equal those of the GPT-2 model. Varying δ from 0 to 1, the resource-rational lossy-context surprisal model thus interpolates between a predictive model without any context, and a full transformer language model.

Implementation based on subwords An important limitation of the original implementation from Hahn et al. (2022) is that it uses a traditional word-based tokenization, with a vocabulary of 50K words. While sufficient to model their experimental stimuli, the model frequently faces OOV tokens when applied to other data, hindering broader validation.[2] In order to apply the model to other experimental stimuli, we straightforwardly adapted the model to modern subword-based tokenizations: Assume a word w consists of tokens $t_1 \ldots t_N$, each represented by token embeddings $e_1 \ldots e_N$, where $N \leq N_{max} = 5$.[3] We concatenate $e_1, \ldots, e_N$ to a vector of length $N \cdot d$ and pad with zeros to obtain a vector of length $N_{max} \cdot d$; we then use a trainable one-layer ReLU MLP to transform this vector into the vector g_i fed into (5), in place of the word embeddings from the original word-based model.[4] When a word x_i has been forgotten, it is represented in y as a single special token, $LOST$, indicating that a word was present but not how many tokens it spanned. Hence, while the model is now specified in terms of subwords, it continues to implement the same cognitive theory; in particular, forgetting continues to apply on the level of words.[5]

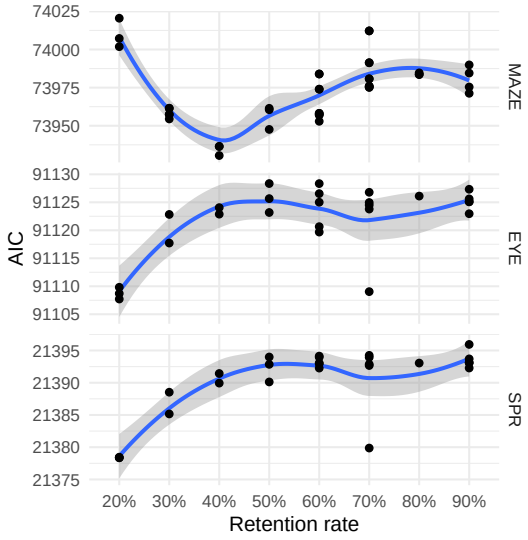

Figure 2: Linear mixed-effects model fit for LCS to Maze (Hahn et al., 2022), ET, and SPR data for filler items from Vasishth et al. (2010). Points are individual LCS model instances, line shows GAM smooth, x-axis shows retention rate, y-axis shows goodness of fit in AIC — lower is better. Maze data are better approximated by LCS with a higher retention rate (40%) compared to ET and SPR data (20%).

Setup We train the model using this parameterization, using the GPT-2 Tokenizer, and otherwise matching the setup of Hahn et al. (2022): The model is trained, separately for different values of δ, on the same English Wikipedia corpus (2.3 billion words). Paragraphs are shuffled and separated by an EOS token. The model is applied to contexts of size T across sentence and paragraph boundaries. In evaluation, the context is padded or truncated to length T (long enough to cover the experimental stimuli); padding is removed before passing to the GPT-2 model. We set $T = 20$.

[2] For example, 8% of the stimuli evaluated in §4 contain at least one OOV under the original model.

[3] In very rare cases of longer words, the tokens starting from the sixth one were disregarded.

[4] In preliminary experiments, we also considered alternative parameterizations, such as simply summing embeddings without any nonlinear transformation. We compared the options at $\delta = 10$, and chose the one with the best result on the objective function (3-4).

[5] Note that another option would be to apply the model at the level of subwords, but this would be of unclear cognitive plausibility, as subwords do not directly correspond to any units of theoretical cogniitive interest, and even depend on tokenizers.

3.3 Evaluation

Hahn et al. (2022) validated that, with a nonzero forgetting rate, their LCS implementation improved fit to Maze RTs on their filler sentences when compared to a model variant with zero forgetting rate. These filler sentences had previously been used in ET and SPR experiments by Vasishth et al. (2010). Crucially, for these fillers, RT data is available from three paradigms: Maze from Hahn et al. (2022), ET and SPR from Vasishth et al. (2010). The fillers comprise both critical items and fillers from Grodner and Gibson (2005, Expt. 1). The items contain a mixture of syntactic structures, including some embedded structures. The key advantage of these filler data compared to datasets such as the Dundee corpus (Kennedy and Pynte, 2005) or Natural stories (Futrell et al., 2017) is that data from *three* paradigms—Maze, SPR, and ET—is publicly available for *exactly the same sentences*, neutralizing confounding effects of factors such as genre.

We evaluate our subword model implementation on the same stimuli and range of modalities. This evaluation has two goals: 1) to confirm that our subword implementation achieves comparable fits to reading time data as the original word-based model, in the sense that relatively low retention rates should model RT better than high retention rates, and 2) to inform our later analysis of task differences in relative clause processing. We model reading time fit per word using the same linear mixed-effects model structure[6] as Hahn et al. (2022, Supp. Mat. §9). We also report goodness of fit in terms of Akaike's An Information Criterion (AIC).

Our findings (Fig. 2) are qualitatively similar to those of Hahn et al. (2022, Supp. Mat. Fig. 30). We observe a comparable spread of AIC values across retention rates, with an average $\Delta AIC \geq 10$ separating the best-fitting retention rate from others. This stark differentiation in goodness of fit suggests that the best-fitting retention rate captures meaningful variation in reading time. Moreover, in line with other literature (§2), we also see that memory constraints — i.e. retention rates much lower than 100%[7] — produce superior fits to human RT data.

We also reproduce the task-specific trends reported by Hahn et al.. They found that Maze RTs were best modeled at a higher retention rate of 5 out of 20 words (25%; compare to 40% in our implementation) compared to ET and SPR RTs, which were best fit at 3 out of 20 words (15%; compare to 20% in our implementation). The remainder of this paper investigates whether these task-dependent differences can account for discrepant empirical results from the relative clause literature.

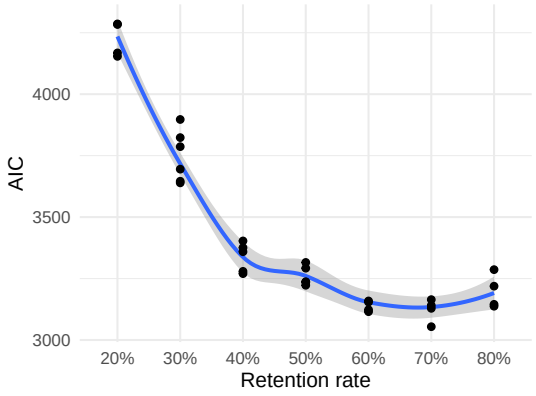

Figure 3: Linear mixed-effects model fit for LCS to Maze RT data on filler items (Vani et al., 2021). Points are individual LCS model instances, line shows GAM smooth, x-axis shows retention rate, y-axis shows goodness of fit in AIC. Retention rate 60–70% achieves the best fit on average.

4 Modeling Relative Clause Processing

The increased difficulty in processing object relative clauses (ORCs) compared to subject relative clauses (SRCs) provides a testing ground for effects of memory and expectation. Memory-based accounts such as Dependency Locality Theory (DLT; Gibson, 1998; Gibson et al., 2000) predict increased reading time (RT) at the ORC verb, reflecting integration of long-distance dependencies. This prediction has been realized in eye-tracking (ET) studies (Traxler et al., 2002; Staub, 2010). The expectation-based Surprisal Theory (Hale, 2001; Levy, 2008), however, predicts an RT slowdown only at the start of the ORC noun phrase, and this pattern has been found in Maze studies (Forster et al., 2009; Vani et al., 2021). Vani et al. suggest that the ORC verb slowdown found in eye-tracking studies may reflect spillover effects rather than memory constraints.

We explore the alternative hypothesis that ET experiments impose lower memory demands relative to the Maze task. At lower retention rates, lossy context surprisal (LCS) models memory con-

[6] LMER formula: $log(RT) \sim LCS + wordPositionInItem + log(WordFreq) + WordLength + prevWordLCS + log(prevWordFreq) + prevWordLength + log(prevWordRT) + (1|ItemID) + (1|ParticipantID)$

[7] Note that LCS with 100% retention rate is functionally equivalent to pure language model surprisal, i.e. GPT2-Medium in our implementation.

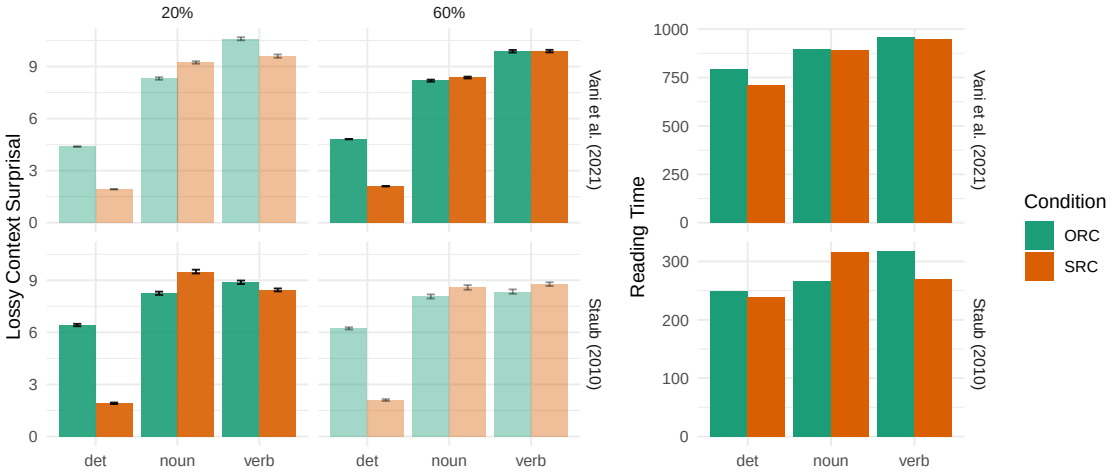

Figure 4: LCS predictions (left; error bars show standard error across model instances and items) and reading time data (right) for stimuli from Staub (2010, ET gaze duration, Experiment 1) and Vani et al. (2021, Maze, Experiment 1; cf. their Figs. 3 and 4). At the higher retention rate (60%), LCS predicts only the determiner slowdown observed in Maze data (top row). At the lower retention rate (20%), LCS also predicts the ORC verb slowdown observed in ET data (bottom row).

straints, but not spillover effects; if LCS captures the patterns in ET behavioral data, this supports the interpretation that ORC verb slowdowns are memory-driven but also modulated by task demands. Using our LCS implementation with subword tokenization, we generate predictions on critical RC stimuli and compare them to behavioral results from Maze (Vani et al., 2021) and eye-tracking (Staub, 2010; Roland et al., 2021). We draw on Roland et al.'s statistical analysis to further distinguish spillover and memory effects.

4.1 Selecting Retention Rate

Maze We use the same evaluation procedure described in §3.3 on the Maze filler item RT data from Experiment 1 of Vani et al. (Fig. 3). Note that these model fits span a broad range of AIC values, so we can confidently state that LCS at higher retention rates better predicts RT data from this experiment. We observe similarly high performance at 60% and 70% retention rates. As the evaluation in §3.3 found a lower retention rate (40%) provided the best fit to Maze data, we conservatively select 60% as more consistent with our earlier analysis.[8]

[8]This difference — 60%–70% retention, vs. the 40% found in §3.3 — may also reflect task demands. Hahn et al. (2022) use the A-Maze task, in which participants distinguish words from length-matched words with low contextual probability. Vani et al. (2021) introduce the I-Maze task variant, which interpolates lexical and grammatical competitors and may impose higher memory demands.

Eye-tracking Unfortunately, filler data is not available for either of the ET studies we aim to model. We select 20% as our prospective retention rate based on the evaluation in §3.3. This low retention rate is consistent with our hypothesis of reduced memory demand in ET studies.

4.2 Evaluating Relative Clause Processing

The previous section identified two distinct retention rates at which to evaluate LCS, based on their fit to reading times from the Maze and eye-tracking experimental settings. In this section, we generate LCS predictions at these two retention rates for the critical relative clause items tested by Vani et al. (2021), Staub (2010), and Roland et al. (2021). Predictions at each retention rate are averaged over multiple LCS model instances trained with different random seeds and hyperparameter configurations, with a minimum of four instances per retention rate. We then compare the predictions to the behavioral patterns reported on these stimuli for Maze and eye-tracking data.

4.2.1 Eye-tracking vs. Maze

We hypothesize that participants systematically engage their working memory at higher capacity during the Maze task compared to the more naturalistic eye-tracking while reading setting. If this is the case, then we expect that LCS at higher retention rates will predict the relative clause processing behavior observed in Maze studies, with an ORC

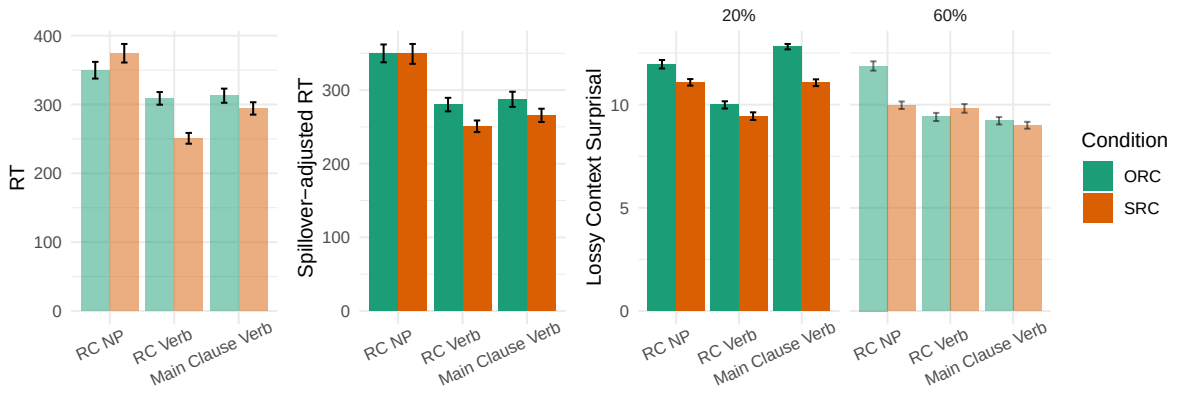

Figure 5: Original and spillover-adjusted gaze duration RT data (left; error bars show standard error across participants and items) and LCS predictions (right; error bars show standard error across model instances and items) for full-NP stimuli from Roland et al. (2021, Experiment 2). At the lower retention rate (20%), LCS predicts ORC slowdowns on the RC verb, consistent with both the original and spillover-adjusted RT data.

slowdown at the beginning of the RC noun phrase, i.e. on the determiner. Conversely, we expect LCS at lower retention rates to predict the pattern of effects observed in eye-tracking studies, with the main ORC slowdown appearing on the RC verb.

LCS predictions largely conform to the expected patterns (Figure 4). At a 60% retention rate, LCS mirrors the processing behavior of participants in Vani et al.'s Maze task, with an ORC slowdown at the determiner but not at the RC noun or verb. At 20% retention, however, we see an ORC slowdown on the RC verb, and a relative ORC speedup on the RC noun — both effects reported in gaze duration ET data from Staub (2010). Crucially, we see the same pattern in LCS predictions across experiments. Vani et al. use test items from Traxler et al. (2002) in their Experiment 1, which differ from the critical items in Experiment 1 of Staub (2010). Nonetheless, the same memory-based pattern — ORC slowdown on the RC verb and speedup on the RC noun — emerges in low-retention LCS predictions for both sets of stimuli.[9]

We use linear mixed-effects models[10] to assess the reliability of these patterns for each retention rate, critical region, and experiment. At the **determiner**, both the high- and low-retention LCS models predict a large and significant ORC slowdown for both experiments. This aligns with the Maze data, but not with the ET data; there is a small ORC slowdown on the determiner, but it is not significant per the statistical analysis of Staub (2010). We speculate that this absence may reflect spillover in the SRC condition, as the determiner directly follows the RC verb; this could raise RT times compared to the ORC condition (in which the determiner follows"that"), obscuring the ORC slowdown effect. At the RC **noun**, both LCS models predict a significant ORC *speed-up*: small at 60% retention, much larger at 20% retention. This appears consistent with the RT data — while Vani et al. report no RC effect here with Maze, Staub finds a significant ORC speed-up in gaze duration. Finally, at the RC **verb**, LCS captures the critical pattern: no ORC slowdown with high retention, as seen in the Maze data — but significant ORC slowdown at low retention, as seen in the ET data. This pattern supports a memory-based rather than spillover interpretation of the ORC verb effect.

4.2.2 Memory vs. Spillover

To further investigate the role of spillover effects in eye-tracking, we draw on the data and analysis of Roland et al. (2021). Their Experiment 2 also compares ORC and SRC processing on a distinct set of RC stimuli.[11] Roland et al. also conduct

[9] We also generate LCS predictions at both retention rates for Experiment 2 from Vani et al./Staub, which served as a control comparison between ORCs and embedded sentence complements in both studies. LCS predictions capture the target effect and do not vary across retention rates — as expected for a control experiment with no predicted memory effects — so we do not consider these findings further.

[10] LMER formula: $LCS \sim Condition + (1|ItemID) + (1|ModelID)$

[11] Roland et al. (2021) include an additional manipulation of NP type, in which the RC noun is either a full noun phrase or a pronoun. For simplicity, we consider only the full NP stimuli here.

an extensive statistical analysis of spillover effects on their gaze duration data. We use the estimated coefficients from their fully specified model (2021, Table 12) to adjust RT values while controlling for spillover.[12]

Recall that the key prediction of memory-based accounts is an ORC slowdown on the RC verb. Figure 5 shows that this effect is visible in the original gaze duration data, and remains after adjusting for spillover. It also shows that this ORC verb slowdown is predicted by LCS at 20% retention, but not at 60% retention — a pattern consistent with the findings of the previous section. Linear mixed-effect model analysis confirms that both high- and low-retention LCS models predict a significant effect of RC type at the verb, but in opposed directions: the 60% retention model predicts an ORC speed-up, while the 20% retention model predicts an ORC slowdown, consistent with the spillover-adjusted RT data. Once again, the observed pattern supports a memory-based account of the RC verb effect observed in ET gaze data.

5 Discussion

Our main finding is that low-retention LCS reproduces key predictions of memory-based accounts, and provides a plausible fit to ET data — whereas high-retention LCS reproduces expectation-based predictions, and better fits Maze data. The Maze task requires that participants actively reject distractor words and select the correct sentence continuation; this activity strikes us as clearly more cognitively demanding than naturalistic reading, so task-dependent memory demands present a viable explanation for these discrepant results.[13] An alternative hypothesis suggested by Vani et al. (2021) attributes the ORC verb slowdown seen in ET data to spillover effects. Our analysis indicates that this is unlikely: the ORC verb slowdown is consistently predicted by low-retention LCS, pointing toward a memory-driven explanation.

To be clear, we do not claim that spillover has *no* systematic influence on relative clause processing. The detailed modeling analysis conducted by Roland et al. (2021) indicates that spillover at least partly accounts for the ORC verb slowdown. The slowdown effect persists, however, even after adjusting for spillover, and our LCS simulations suggest that the slowdown reflects memory constraints (Figure 5).

We note that LCS consistently predicts some patterns which have not been given a formal theoretical articulation. Further investigation is required to assess when these discrepancies could be systematic and theoretically meaningful. The ORC noun speed-up presents an interesting case study: this effect is not directly predicted by either expectation or memory accounts, but it appears robustly in both LCS predictions and the ET data for Experiment 1 of Staub (2010). This unexpected concordance suggests that memory constraints may also drive this effect. On the other hand, LCS appears to incorrectly predict an ORC slowdown at the RC NP for the Roland et al. (2021) stimuli (Figure 5); however, closer analysis reveals that this effect is driven by the ORC slowdown at the determiner — on the RC noun itself, LCS once again predicts an ORC speedup, and this effect is larger at the lower retention rate of 20%.[14] Under LCS, memory constraints appear to drive both the ORC verb slowdown and the ORC noun speedup, although to our knowledge the latter effect has not been discussed in connection with memory-based accounts. Exploring the nature of this connection could be a promising direction for future research.

Future work could also explore alternative approaches to modeling expectation. While surprisal theory is well-represented in the research literature and closely aligned with the standard language model learning objective, other research has formulated expectation in terms of *information gain* (e.g. Hale, 2016; Hoover, 2024). Under an information gain account, the incremental cost of processing a given word reflects not its conditional probability (as posited by surprisal theory), but rather the *uncertainty reduction* it provides between alternative sentence continuations. Chen and Hale (2021) use one such approach, namely Entropy Reduction (Hale, 2003), to model the same relative clause processing asymmetry addressed here. They use corpus statistics to compute word-by-word transitions in entropy over the probabilities of following syntactic derivations, and find that this measure

[12] Note that we adjust only for spillover predictors, not for other estimated effects.

[13] While tasks with higher cognitive load are often associated with *reduced* memory capacity in the research literature, we note that the cognitive load in the Maze task is not opposed to sentence processing, but in fact perfectly aligned with it. Higher retention of the preceding sentence context will facilitate higher performance on the task itself, i.e. selecting the correct sentence continuation.

[14] We are unable to compare this prediction directly to the Roland et al. ET data, as RTs are reported for critical regions rather than individual words.

predicts the observed ORC slowdown at both the RC NP determiner and the RC verb. Their model can therefore account for the ORC verb slowdown observed in ET data — however, it would not appear to predict the pattern observed in Maze data by Vani et al. (2021). An alternative information gain approach (e.g. Hoover, 2024) could in principle address such task-dependent effects. In the meantime, we note that LCS straightforwardly captures this variation in relative clause processing as a consequence of memory demands.

Other avenues for future research could address further limitations of the current study. For instance, it might be more appropriate to vary retention rates not only at the experiment level, but also to model differences between individual participants. One could also pursue more interpretability in LCS predictions through detailed analysis of specific word-level reconstructions. Lastly, this paper focuses on one grammatical phenomenon in one language; a thorough treatment of memory effects in online language comprehension will naturally require a broader scope of evaluation.

6 Conclusion

We find that manipulating the retention rate of a lossy context surprisal (LCS) model captures task-dependent differences observed in reading times (RTs). Filler item RTs from the Maze task are best fit with a relatively high retention rate (e.g. 60%), while lower retention (20%) better predicts eye-tracking RTs for those same items. Furthermore, based on these task-dependent retention rates, LCS correctly predicts critical RT patterns observed for English relative clauses. In particular, low-retention (20%) LCS follows memory-based theories and predicts higher RTs for object relative verbs — an effect found in eye-tracking but not Maze studies. These results can explain the apparently contradictory behavioral evidence supporting both memory- and expectation-driven accounts: relative clause processing is likely modulated by the memory demands of the task, and we can use LCS to model this phenomenon.

Acknowledgments

The authors are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

John R Anderson and Robert Milson. 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703.

John R Anderson and Lael J Schooler. 1991. Reflections of the environment in memory. *Psychological science*, 2(6):396–408.

Brian Bartek, Richard L. Lewis, Shravan Vasishth, and Mason R. Smith. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178–1198.

Veronica Boyce, Richard Futrell, and Roger P. Levy. 2020. Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:104082.

Veronica Boyce and Roger Levy. 2020. A-maze of natural stories: Texts are comprehensible using the maze task. In *Talk at 26th Architectures and Mechanisms for Language Processing conference (AMLaP 26). Potsdam, Germany*.

Zhong Chen and John T. Hale. 2021. Quantifying Structural and Non-structural Expectations in Relative Clause Processing. *Cognitive Science*, 45(1):e12927.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Kenneth I. Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171.

Felicity F. Frinsel and Morten H. Christiansen. 2024. Capturing individual differences in sentence processing: How reliable is the self-paced reading task? *Behavior Research Methods*.

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

Richard Futrell, Edward Gibson, Hal Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. 2017. The natural stories corpus. *arXiv preprint arXiv:1708.05763*.

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

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Daniel Grodner and Edward Gibson. 2005. Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity. *Cognitive Science*, 29(2):261–290.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119. Publisher: Proceedings of the National Academy of Sciences.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

John Hale. 2003. The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*, 32(2):101–123.

John Hale. 2016. Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, 10(9):397–412.

Jacob Louis Hoover. 2024. *The cost of information: Looking beyond predictability in language processing*. Ph.D. thesis, McGill University.

Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision research*, 45(2):153–168.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context Limitations Make Neural Language Models More Human-Like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Roger Levy. 2013. Memory and surprisal in human sentence comprehension. In *Sentence processing*, pages 78–114. Psychology Press.

Richard L. Lewis and Shravan Vasishth. 2005. An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, 29(3):375–419.

Byung-Doh Oh and William Schuler. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422. Place: US Publisher: American Psychological Association.

Douglas Roland, Frederic Dick, and Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379.

Douglas Roland, Gail Mauner, and Yuki Hirose. 2021. The processing of pronominal relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 119:104244.

Adrian Staub. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

William Timkey and Tal Linzen. 2023. A Language Model with Limited Memory Capacity Captures Interference in Human Sentence Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Matthew J Traxler, Robin K Morris, and Rachel E Seely. 2002. Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language*, 47(1):69–90.

Pranali Vani, Ethan Gotlieb Wilcox, and Roger Levy. 2021. Using the Interpolated Maze Task to Assess Incremental Processing in English Relative Clauses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *arXiv preprint*. ArXiv:2307.03667 [cs].

Naoko Witzel, Jeffrey Witzel, and Kenneth Forster. 2012. Comparisons of Online Reading Paradigms: Eye Tracking, Moving-Window, and Maze. *Journal of Psycholinguistic Research*, 41(2):105–128.