Title
How Do Syntactic Statistics and Semantic Plausibility Modulate Local Coherence Effects

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Hao, Hailin
Hahn, Michael
Kaiser, Elsi

Publication Date
2023

How Do Syntactic Statistics and Semantic Plausibility Modulate Local Coherence Effects

Hailin Hao (hailinha@usc.edu)
Department of Linguistics, University of Southern California
Los Angeles, CA 90089-1693 USA

Michael Hahn (mhahn@lst.uni-saarland.de)
Department of Language Science and Technology, Saarland University
66123 Saarbruecken, Germany

Elsi Kaiser (emkaiser@usc.edu)
Department of Linguistics, University of Southern California
Los Angeles, CA 90089-1693 USA

Abstract

Local coherence is a phenomenon in human sentence processing whereby word sequences within a sentence incur processing difficulty when they have a plausible reading different from their true syntactic structure as disambiguated by the global context. Prior research (Tabor, Galantucci, & Richardson, 2003) indicates that more plausible substrings incur more processing difficulty than less plausible ones. In the current article, we challenge this view by providing evidence from two experiments which show that local semantic plausibility can actually facilitate processing. We additionally test whether syntactic statistics can modulate local coherence effects, a prediction made by Lossy-Context Surprisal (LCS; Futrell, Levy, & Gibson, 2020; Hahn, Futrell, Levy, & Gibson, 2022). Although we do not find evidence for effects of syntactic statistics, our overall results cannot be fully explained by any existing account of local coherence alone. We discuss implications for theories of sentence processing.

Keywords: local coherence; sentence processing; surprisal; cue-based retrieval

Introduction

The human sentence parser can sometimes be misled by a substring that has a locally coherent reading which is globally impossible within the preceding context. In a self-paced reading study, Tabor, Galantucci, and Richardson (2003) found that the critical verb *tossed* incurs higher processing effort in (1a) in comparison to *tossed* in (1b), or *thrown* in (1c). The key difference, they argued, is that in (1a), the substring "the player tossed a frisbee" has a locally coherent reading as a clause, which is in conflict with the globally correct reading required by the preceding context. In contrast, the corresponding strings in (1b-c) ("the player who was tossed a frisbee", "the player thrown a frisbee") do not have such a reading available. They argued that the effects are best explained by a self-organized parsing model (SOPARSE), where lexical items bring with them small pieces of syntactic structure that may freely combine at any point during processing and attempt to form a globally well-formed tree structure. Processing difficulty arises when multiple locally well-formed attachments compete with each other.

(1) a. The coach smiled at *the player tossed a frisbee* by the opposing team.
b. The coach smiled at *the player who was tossed a frisbee* by the opposing team.
c. The coach smiled at *the player thrown a frisbee* by the opposing team.

Tabor et al. (2003) also show that when a substring is semantically more plausible, it triggers more processing difficulty. In the following example, (2a) is more difficult to process at the critical verb *transported* than (2b), because in the former, a subject-verb parse of the local string "prisoner transported" is semantically plausible, while a subject-verb parse of the local string in (2b; "gold transported") is implausible. According to SOPARSE, this is due to the fact that semantically plausible local strings cause more competition or interference with the global analysis.

(2) a. The bandit worried about the *prisoner transported* by the capricious guards.
b. The bandit worried about the *gold transported* by the capricious guards.

A rational belief-update account formalizing similar intuitions is provided by Bicknell and Levy (2009). A different account of local coherence is provided by the Noisy-Channel Error-Detection model (Levy, Bicknell, Slattery, & Rayner, 2009), which assumes that perceptual input from a given word can give rise to multiple potential word representations. For example, in the case of (1a), when a comprehender encounters the local strings "the player tossed a frisbee", they might revise their previous input as "the coach smiled *as*" (instead of *at*) to make the later input more congruent. Under this view, however, the revision of past input, however, is costly. This is reflected by longer processing time or more regressions in eye movement during reading reported in Levy et al. (2009).

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Open Questions and Aims of this Work

Since their discovery, local coherence effects have become an important testing ground for theories of sentence processing (Bicknell & Levy, 2009; Paape, Vasishth, & Engbert, 2021; Cutter, Filik, & Paterson, 2022). Over the past decades, many aspects of syntactic processing difficulty have been unified in broadly successful accounts based on expectation violation (Hale, 2001; Levy, 2008) or memory retrieval (e.g., Lewis & Vasishth, 2006), but local coherence has so far withstood attempts at unification with those theories, making it particularly interesting as a missing piece in the larger theoretical picture. As a consequence, there is still no consensus regarding the underlying cognitive mechanisms of local coherence effects.

While the local coherence effect in (1a) has been replicated in multiple studies (Levy et al., 2009; Christianson et al 2017; Cutter et al., 2022), the broader occurrence of this phenomenon, and broader predictions of the existing accounts, remain underexplored, and a null effect in a different configuration is reported by Kauf and Levy (2022). Understanding the broader distribution of local coherence effects is critical for understanding how it can be integrated into the larger theoretical picture, and a prerequisite to developing unified theories of sentence processing.

In the present paper, we explore how semantic and statistical cues modulate local coherence effects and test the predictions of Lossy-Context Surprisal (LCS; Futrell, Levy, & Gibson, 2020; Hahn, Futrell, Levy, & Gibson, 2022) against existing accounts SOPARSE (Tabor et al., 2003), the belief update model (Bicknell & Levy, 2009) and the Noisy-Channel Error-Detection model (Levy et al., 2009). In the remainder of this section, we detail the architecture of LCS and what unique predictions it makes for local coherence effects, in comparison to exitising accounts.

Lossy-Context Surprisal and Its Predictions

LCS is built on Surprisal (Hale, 2001; Levy, 2008), a general psycholinguistic theory arguing that the processing difficulty of a word is proportional to its contextual predictability, as formulated in (2).

(2) processing difficulty $\propto -\log \mathrm{P}(w|c)$

The standard Surprisal theory, however, requires the context to be perfectly retained in comprehender's memory. By contrast, LCS acknowledges that processing is not only expectation-based, but also constrained by memory (Gibson, 1998; Lewis & Vasishth, 2005) and noisy-channel inferences (Gibson et al., 2013). For rational comprehenders, when memory representations of context are not perfect, they can be reconstructed based on the statistics of the language. Therefore, next-word processing difficulty is not only determined by the veridical input, but also by variants with high *a priori* probability that are similar to the true context. The idea is formulated in (3), where next word predictability is determined by a posterior $\mathrm{P}(c|c')$ over possible contexts calculated via Bayes' rule and knowledge of the *a priori* language statistics. However, the original formulation of LCS leaves open which aspects of previous input are prone to noisy-channel edits, which makes it hard to derive testable predictions.

Hahn et al. (2022) offered a computational implementation of LCS, using a resource-rational model (Lieder & Griffiths, 2019) of fine-grained memory representation, which proposes that rational agents with limited cognitive resources should optimize their memory so as to minimize expected downstream processing effort. This model, Resource-Rational LCS (RR-LCS), calculates a retention probability for each past word, which is optimized to minimize the average model surprisal calculated from GPT-2 (Radford et al., 2019) over large-scale text data from the English Wikipedia. Model outcomes suggest that words that are less recently encountered and more frequent are more likely to be forgotten.

One key prediction of the model in Hahn et al. (2022) concerns a-prior syntactic statistics in the processing of recursive structure. Specifically, they investigated *embedding bias*, the statistical tendency for a noun like "fact" or "report" to be followed by a complement clause such as in "the fact/report that the prisoner escaped is worrying". They show that with head nouns whose embedding bias is high (e.g., "fact"), recovering from an enbedded complement clause is easier than in the case of head nouns whose embedding bias is low (e.g., "report"). This is attributed to the embedded structure being represented more faithfully in working memory when the noun's statistical properties support this structure.

(3) $-\log P(w|c') = -\log \sum_c P(w|c)\, P(c|c')$

Predictions of LCS and Other Existing Accounts

For local coherence, RR-LCS makes the prediction that semantically more plausible local strings can sometimes be easier to process than less plausible counterparts. In Example (4), "the surgeon/officer has arrested" has a locally coherent reading that is inconsistent with the only possible global analysis, where the subject of the verb phrase "has arrested" is the head noun "report/fact", which is semantically implausible. However, according to RR-LCS, a comprehender's working memory might sometimes fail to retain a representation of the function word *of* and subsequently reconstruct it as *that.* If this happens, the local string "the surgeon/officer has arrested" becomes consistent with the global analysis required by the comprehender's representation of past input, as in "the report/fact that the surgeon/officer has arrested the thief" (note that when this happens, (4) will be analyzed as an NP fragment, instead of a complete SVO structure). When the locally coherent substring is more plausible (e.g., "the officer has arrested"), the surprisal at the critical verb *arrested* should be lower than when the substring is less plausible (e.g., "the surgeon has arrested"). In addition, RR-LCS predicts that the facilitating effect of plausibility effect should be stronger when the head noun has high embedding bias (e.g., "fact"), since the chance of misrepresenting *of* as *that* depends on the noun's embedding bias.

(4) The report/fact of the surgeon$_{IMPLAUS}$ / officer$_{PLAUS}$ has arrested the thief.

A facilitation due to increased local plausibility is also predicted by cue-based retrieval (Lewis & Vasishth, 2005) if we consider the head noun as the target and the substring subject as the distractor. Related effects are indeed reported in Cunnings and Sturt (2018) and Laurinavichyute and von der Malsburg (2022), though not in online processing in the presence of local coherence.

Having considered LCS, we now consider the predictions of the existing accounts of local coherence. For SOPARSE, plausible substrings will incur greater competition between local and global analyses, hence more processing difficulty. The same holds for the belief-update account (Bicknell & Levy, 2009). The Noisy-Channel Error-Detection theory also appears to predict that plausible substrings require more processing efforts because they are more likely to lead comprehenders to revise their beliefs regarding previous input (e.g., from "report of the officer has arrested" to "report that the officer has arrested"), though the exact predictions will depend on the noise model, for which currently no broadly applicable implementation exists.

In Experiment 1, we test the online predictions of these three theories using A-MAZE (Boyce, Futrell, & Levy, 2020). Experiment 2 aims to validate the results of Experiment 1 using an offline pseudo-production method that asks participants to judge whether the critical verb (e.g "arrested") is a good continuation given the context.

Experiment 1: Maze

Methods

We adopt the A-Maze paradigm (Boyce, Futrell, & Levy, 2020), which records word-by-word Reaction Times (RTs) by making participants choose between the correct next word and a distractor, as illustrated in Fig 1 (selections are marked by blue ovals). If they make a mistake, participants are prompted by an error message to try again and continue with the sentence. The A-Maze leverages neural networks (Gulordava et al., 2018) to automate distractor selection by generating a word that has very low contextual predictability yet matches the correct next work in terms of frequency and length. Maze tasks have been found to provide more localized measures than self-paced reading and eye-tracking (Forster, Guerrera, & Elliot, 2009; Witzel, Witzel, & Forster 2012; Boyce, Futrell, & Levy, 2020).

Moreover, we adopted a one-trial design, meaning that each participant is only presented with one critical trial. This is to avoid task adaptation and the influence of implausible stimuli on overall task performance (e.g., Gibson et al., 2013).

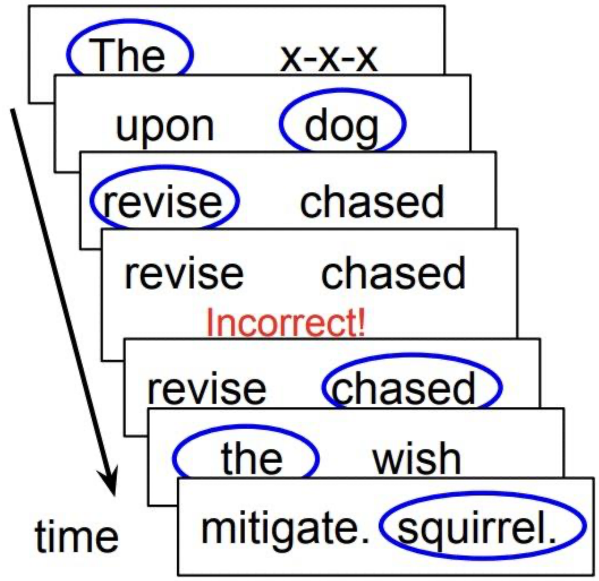

Fig 1: Schematic of the Maze task

Participants 1,198 participants were recruited through the online recruitment platform Prolific (2014). All participants have an IP address in the United States. Participants who self-identify as non-native English speakers are excluded (N=55).

Stimuli We selected 16 nouns with varying levels of embedding bias (e.g., "report" is followed by complement clauses less frequently than "fact", i.e. has lower embedding bias) using the English Wikipedia corpus. We crossed nouns that have relatively high vs. low embedding bias with the binary factor plausibility (high vs. low). Here, plausibility refers to whether the first animate noun is a plausible subject of the verb, "the surgeon has arrested vs. the officer has arrested"). We created 512 stimuli in the form shown in (4). The head noun (e.g., "report/fact") is never compatible with the critical verb (e.g., "has arrested"). The compatibility between the nouns and the verbs was normed using GPT-2 Large (Radford et al., 2019). In all of our stimuli, we made sure that the surprisal value at the verb "arrested" is significantly higher when it is preceded by the head nouns (e.g., "the report/fact") or an implausible animate noun (e.g., "the surgeon") than when it is preceded with a plausible animate noun (e.g., "the officer"). No stimuli contain "the NOUN that"; instead, they all start with "the NOUN of". Our analyses focus on the RT at the critical verb (e.g., "arrested").

Procedure The experiment was conducted online using PCIbex (Zehr & Schwarz, 2018). Participants were first given instructions regarding how Maze works and told that they would read 3-4 sentences or sentence fragments this way. (Participants were told that they might encounter sentence fragments because, as we mentioned earlier, if participants indeed misrepresent the critical stimuli like "The fact of the officer has arrested the thief" as "The fact that the officer has arrested the thief", the stimulus will be rendered as a complex NP phrase instead of a complete sentence). The first two sentences were fillers; the third one was a critical trial. The fillers serve the purpose of familiarizing participants with the task. The experiment took around 2 minutes to complete, and participants received 0.60 US dollars**.**

Analysis

Exclusion Participants with a Maze choice accuracy under 80% for the filler trials and those who made mistakes on the critical trial were excluded (N=179). Trials with RTs at the critical verb over 5000 ms or under 100 ms were also excluded (N=24). The exclusion criteria largely follow Boyce and Levy (2022). We are left with 840 participants for analysis (73.49% of original data). Less conservative exclusion criteria were attempted as well and did not affect conclusions.

Model Structure For statistical analysis, we fitted Bayesian linear mixed effects regression models on log-transformed RTs at the critical word region using the brms package, version 2.12 (Bürkner, 2017) in R. The region of interest is the critical verb. No spillover regions are analyzed since Maze is known to provide localized effects (Forster, Guerrera, & Elliot, 2009; Witzel, Witzel, & Forster 2012; Boyce, Futrell, & Levy 2020). The continuous predictor *embedding bias* is centered relative to the mean of all 16 nouns used in the study. The binary predictor *plausibility* is sum-coded as (-1, 1). The model also includes by-item intercepts and slopes for all fixed effects and their interaction and by-noun intercepts and slopes for *plausibility*. No by-participant intercepts or slopes are included since each participant only saw one critical trial. Relatively uninformative priors are chosen based on previous research (Hahn et al., 2022; Boyce & Levy, 2022), which allows for a plausible yet wide range of Maze RT and effect sizes in either direction.

Results

We first investigated the predictions of RR-LCS. We note that no broad-coverage implementations are available for SOPARSE, cue-based retrieval, or the error-detection account. Predictions from RR-LCS for the experimental materials are shown in Fig. 2.

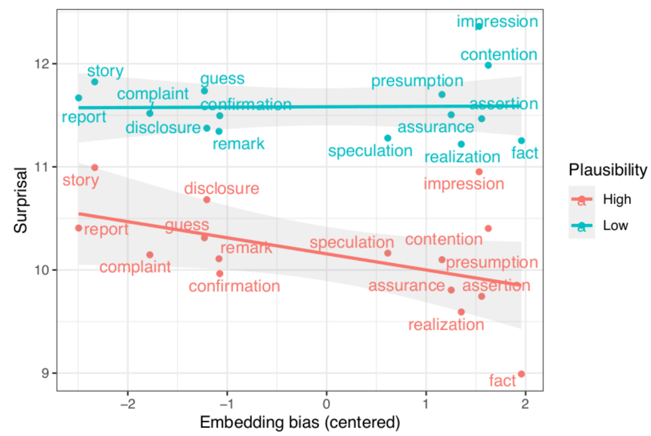

Fig 2: Exp 1 Surprisal predictions for the critical region from RR-LCS (grouped by the noun used and the local plausibility)

We analyzed the simulated results using the same method as for the human data, with by-noun and by-item random effects. Figure 5A has posteriors for fixed effect βs. A main effect of plausibility on surprisal is highly prominent (β=-1.46, CrI=[-1.90,-1.03]), suggesting that plausible substrings reduce lossy-context surprisal as we had conceptually argued in the introduction. There is no main effect of embedding bias (β=-0.08, CrI=[-0.25,0.09]), but there is evidence for a small interaction effect (β=-0.17, CrI = [-0.30,-0.04]), such that the effect of plausibility is stronger on fact-like nouns than on report-like nouns.

The human data from our experiment are in Fig 3; Figure 5B has posteriors for fixed effect βs. Here, we report the results on the log ms scale with 95% credible intervals. The 95% credible interval gives a range of values that contains plausible values of a parameter with 95% probability given the data and statistical model. A main effect of plausibility on RTs is detected (β=-0.071, CrI=[-0.114, -0.028]), suggesting that plausible substrings indeed ease processing. However, there is little evidence for effects of embedding bias (β=0.004, CrI=[-0.020, 0.028]) or for an interaction between the two factors (β=0.004, CrI=[-0.012, 0.034]).

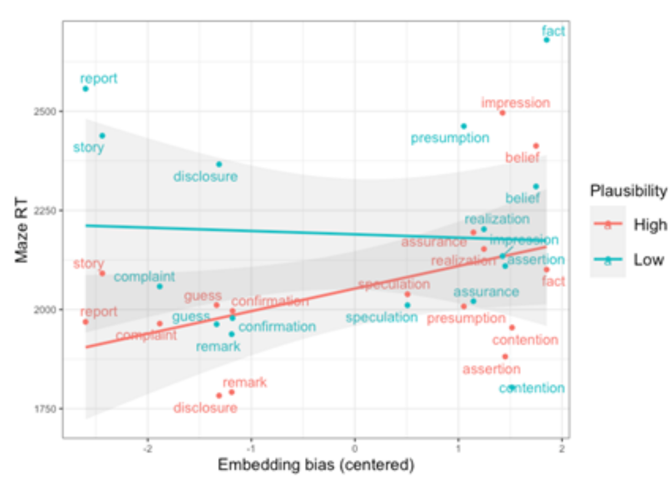

Fig 3: Exp 1 Maze RT for the critical region (RTs in milliseconds; grouped by the noun used and the local plausibility)

Interim Discussion

Echoing RR-LCS, we find facilitative effects of plausibility on the processing of LC, which is not predicted by SOPARSE, error-detection theory or standard Surprisal Theory. However, an interaction with embedding bias predicted by RR-LCS was not confirmed in the human data. To validate the results of this reaction-time study, we also conducted an offline judgment experiment to examine the acceptability of the critical verb given the context:

Experiment 2: Continuation Verification

Methods and Predictions

We adopt a one-trial design using a sentence continuation verification task similar to Laurinavichyute and von der Malsburg (2022). Participants are asked to judge whether, for example, the word *arrested* is a good continuation of a sentence fragment presented to them earlier (e.g., "the report of the surgeon has"). The correct answer should always be *no* here since for the veridicial version of the fragment, the grammatical subject "report" is semantically implausible as the agent of *arrested*.

We predict that accuracy rates should be positively correlated with processing difficulty (critical verbs that are less contextually predictable should be more difficult to process online, as indexed by longer RTs, and incur a higher ratio of "no" answers). Therefore, RR-LCS predicts that the accuracy rates for this task should be lower (i.e., participants should be more likely to accept "arrested" as a good continuation of "the report of the surgeon has") when the local strings are semantically plausible, while SOPARSE and the Noisy-Channel Error-Detection theory predicts that locally plausible strings should incur higher accuracy. RR-LCS in addition predicts that the negative influence of plausibility on accuracy rates should be stronger when embedding bias is high. Same as in Experiment 1, each participant only saw one critical trial.

Participants 1,032 participants were recruited through the online recruitment platform Prolific (2014). All participants have an IP address in the United States. Participants who self-identify as non-native English speakers are excluded (N=42).

Stimuli The same stimuli were used as in Experiment 1, with *embedding bias* crossed with *substring plausibility*.

Procedures The experiment was conducted online using PCIbex (Zehr, & Schwarz, 2018). Participants were told that they would see a sentence fragment first, and then judge whether a word is a good continuation of that fragment. Participants were free to decide for how long they want to read the fragment; there was no time pressure. Once they are ready, they click *continue* and see, on the next page, a question such as "Is *arrested* a good continuation of the fragment you just saw?". Participants answer *yes* or *no* by pressing the F or J key.

Analysis

Exclusion Since the task is making a binary choice, no data points were excluded other than those from non-native English speakers, as mentioned above.

Model Structure For statistical analysis, we fitted Bayesian logistic regression models with a bernoulli link function on accuracy using the brms package, version 2.12 (Bürkner, 2017) in R. The structure of the statistical model is the same as the one used in Experiment 1, with the accuracy rate as the response variable. As regularizing priors for the fixed effects, we follow Laurinavichyute and von der Malsburg (2022) by using a normal distribution with mean 0 and standard deviation 1, which allows for a wide range for effect sizes from 0% up to 90% and discourage only implausibly large effects bigger than 90%.

Results

The results are presented in Fig 4; Figure 5C has posteriors for fixed effect βs. Here, we report the results with 95% credible intervals, which gives a range of values that contains plausible values of a parameter with 95% probability given the data and statistical model. A main effect of plausibility on accuracy is detected (β=-0.714, CrI=[-0.972, 0.456]), suggesting that the plausible substrings are more likely to be accepted. In other words, people are more likely to say *yes* when they see "arrested" after "the NOUN of the officer" than after "the NOUN of the surgeon". However, there is little evidence for a main effect of embedding bias (β=-0.007, CrI=[-0.123, 0.109]) or an interaction between the two predictors (β=-0.052, CrI=[-0.175, 0.069]).

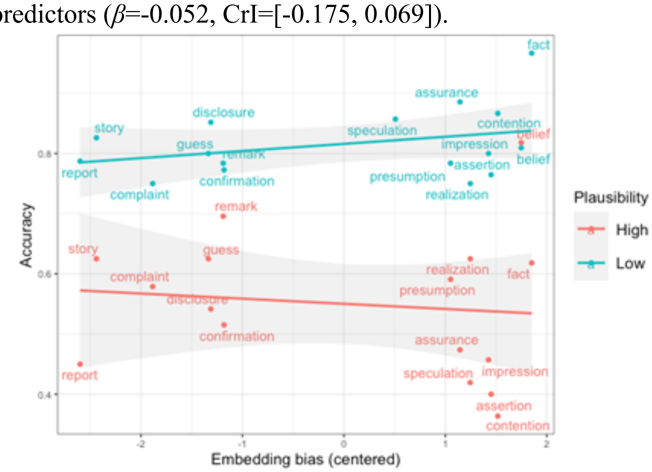

Fig 4: Exp 2 Proportion of accurate answers (i.e., answering *no* in the verification task; grouped by the noun used and the local plausibility)

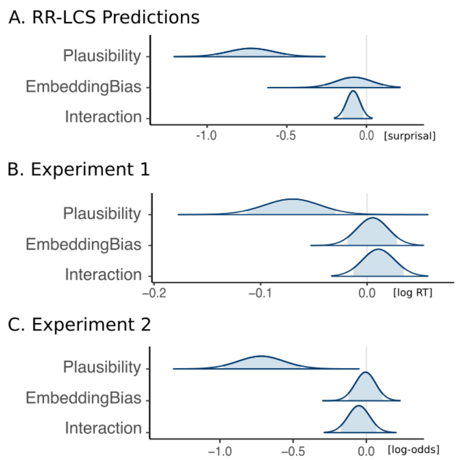

Fig 5: Posteriors for β coefficients across analyses of model predictions and experiments. Model predictions and both experiments show a strong effect of plausibility. A small interaction with embedding bias, also found in the model predictions, is not detectable in the experiments.

General Discussion

In this study, we presented two experiments demonstrating that semantically plausible substrings can *facilitate* processing, contrary to the predictions of the two most prominent accounts of local coherence, SOPARSE and Noisy-Channel Error-Detection. This effect is predicted by RR-LCS (Figure 5), but not by existing accounts of local coherence. It is also conceptually predicted by cue-based retrieval accounts, though the absence of a broad-coverage implementation of cue-based retrieval prevented us from obtaining quantitative model predictions.

RR-LCS additionally predicts a small interaction with embedding bias, but this was not detected in the experiments. Comparing Figures 5A and 5B-C shows that this interaction may be too small to be measured in the experiments, which neither confirm nor disconfirm the presence of an interaction of the predicted size (about 5% of the plausibility main effect β). The small predicted size of this interaction, compared to the findings in Hahn et al (2022), might reflect an edit asymmetry in noisy channel inferences, given that adding a word is assumed to be less likely than deleting one (Gibson et al., 2013): In Hahn et al. (2022), where the embedding bias was found to modulate the processing of recursive structures, the processing difficulty is linked to how likely *that* is retained, whereas in in studies, *that* would have to be added.

The present work adds to research on local coherence phenomenon by demonstrating that increased local plausibility can facilitate, rather than inhibit, processing. This result is incompatible with the existing theoretical accounts of local coherence but is predicted by LCS and cue-based retrieval. These accounts, on the other hand, do not predict the classical local coherence effect in (1). This raises the question of what theoretical account can explain the full range of local coherence effects. One possibility is that a unified model would need to incorporate both groups of theories, and the facilitative effects predicted by LCS and cue-based retrieval override the processing disruption traditionally associated with locally coherent structures. Another possibility is that accounts of local coherence overpredict its occurrence, and that this disruption simply does not occur in the configurations we investigate in this paper. Such a possibility is also suggested by a null effect reported by Kauf and Levy (2022) for a different structure, in which SOPARSE may likely predict difficulty induced by local coherence. Developing unified accounts of syntactic processing difficulty that can predict the correct distribution of effects is an important problem for future research; our experiments provide data towards this goal.

Acknowledgments

M.H. acknowledges support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID 232722074 — SFB/CRC 1102.

References

Bicknell, K., & Levy, R. (2009). A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. In North American Chapter of the Association for Computational Linguistics.

Boyce, V., & Levy, R. (2022). A-maze of natural stories: Comprehension and Surprisal in the maze task. https://doi.org/10.31234/osf.io/8xkrf

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language, 111*, 1-13.

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

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405.

Cunnings, I., & Sturt, P. (2018). Retrieval interference and semantic interpretation. *Journal of Memory and Language, 102*, 16–27.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods, 41* (1), 163–171.

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

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).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 conference of the North American chapter of the association for* *computational linguistics: Human language technologies* (pp. 1195–1205).

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*.

Kauf, C. and Levy, R. (2022). No local coherence effects in the absence of high surprisal. In *35th Annual Conference on Human Sentence Processing*.

Laurinavichyute, A., & von der Malsburg, T. (2022). Semantic attraction in sentence comprehension. *Cognitive Science*, *46*(2).

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

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

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29* (3), 375–419.

Paape, D., Vasishth, S., & Engbert, R. (2021). Does local coherence lead to targeted regressions and illusions of grammaticality? *Open Mind*, 1–17.

Prolific. (2014). https://www.prolific.com

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

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language, 50* (4), 355–370.

Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research, 41* (2), 105–128.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832