Can Syntactic Log-Odds Ratio Predict Acceptability and Satiation?

Jiayi Lu*, **Jonathan Merchan***, **Lian Wang***, **Judith Degen**
Department of Linguistics, Stanford University
{jiayi.lu, jmerchan, lianwang, jdegen}@stanford.edu

Abstract

The syntactic log-odds ratio (SLOR), a surprisal-based measure estimated from pre-trained language models (LMs) has been proposed as a linking function for human sentence acceptability judgments, a widespread measure of linguistic knowledge in experimental linguistics. We test this proposal in three steps: by examining whether SLOR values estimated by GPT-2 Small predict human acceptability judgments; by asking whether satiation effects observed in human judgments are also exhibited by fine-tuned LMs; and by testing whether satiation effects generalize in qualitatively similar ways in the model compared to humans. We show that SLOR in general predicts acceptability, but there is a significant amount of variance in acceptability data that SLOR fails to capture. SLOR also fails to capture certain patterns of satiation and generalization. Our results challenge the idea that surprisal alone, via a SLOR linking function, constitutes an accurate model for human acceptability judgments.

1 Introduction

Judgments of a sentence's acceptability are commonly interpreted as a reflection of linguistic knowledge. For example, native English speakers find sentences like **What did John hear the rumor that Mary ate?* much less acceptable than sentences like *What did John hear that Mary ate?*. These kinds of acceptability judgments by native speakers have been widely used to inform linguistic theories. For example, based on the acceptability contrast in the aforementioned two sentences, linguists have proposed syntactic constraints (in this case, the "complex-NP island constraint") to rule out the first sentence as ungrammatical (Ross, 1967).

Despite the widespread use of acceptability judgments as a source of evidence to inform linguistic theories, the cognitive mechanisms involved in generating these judgments are rather poorly understood (Schütze, 1996; Sprouse, 2018; Francis, 2022). Past linguistic research has identified various factors that affect a sentence's acceptability, including but not limited to its grammaticality, the frequency of observed lexical items and structures, task-related factors such as presentation order, and subject-related factors such as literacy and prior linguistic training (Schütze, 1996). However, there is no clearly spelled-out model that captures how these factors interact to give rise to an acceptability judgment. More recently, some studies hypothesized that there is a "surprisal bottleneck" for acceptability judgments: just as surprisal serves as a causal bottleneck for online processing difficulty (Levy, 2008), surprisal may also be the causal bottleneck for sentence acceptability (Lau et al., 2017, 2020; Culicover et al., 2022). If pre-trained language models (LMs) capture human linguistic knowledge, some studies argue that surprisal-based metrics estimated by LMs may serve as linking functions for human sentence acceptability judgments (Lau et al., 2017, 2020). In one prominent study, human sentence acceptability judgments were found to be best predicted by the syntactic log-odds ratio (SLOR, shown in Equation 1) values, a sentence's model-given log probability normalized by its length and its probability based on its lexical items' unigram probabilities (Lau et al., 2017): [1]

$$\text{SLOR} = \frac{\log p_m(s) - \sum_{w \in s} \log p_u(w)}{|s|} \quad (1)$$

Here, $p_m(s)$ is the probability of a sentence s as estimated by the model (calculated as the product of the model-estimated transition probability for each word), $p_u(w)$ is the unigram probability of a word w in s, and $|s|$ is the sentence's length in words. SLOR achieved the best correlation with

*These authors contributed equally.

[1] SLOR was first proposed by Pauls and Klein (2012) for a different task.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 10-19.
Irvine, California, June 27-29, 2024

sentence acceptability ratings among a variety of surprisal-based metrics.

In the present study, we revisit the hypothesis that SLOR estimates from pre-trained LMs provide a good linking function for acceptability judgments. We do so in three ways: first, we replicate the correlation between SLOR and sentence acceptability ratings using a more up-to-date LM than that used by Lau et al. (2017). Second, we move beyond one-shot acceptability ratings and investigate whether the changes in SLOR after fine-tuning predict the changes in human acceptability judgments in response to exposure (i.e. the "satiation effect": Snyder, 2000; Chaves and Dery, 2019; Lu et al., 2021, inter alia). Third, we test whether fine-tuning the model with one sentence type leads to SLOR increase in a different but structurally related sentence type, replicating the generalization of satiation effects found in human acceptability judgment data (Lu et al., 2022).[2]

If the pre-trained LM approximates human linguistic knowledge and its SLOR estimates constitute a good linking function for human sentence acceptability judgments, SLOR values should correlate with acceptability judgments and demonstrate both human-like satiation effects and the generalization of satiation effects – both of which are phenomena that have been shown to reliably emerge in human acceptability judgment tasks (Snyder, 2000; Chaves and Dery, 2019; Lu et al., 2021, 2022).

2 Experiment 1: SLOR Predicts Acceptability

Experiment 1 aims to replicate Lau et al. (2017)'s finding that SLOR predicts sentence acceptability judgments using GPT-2 Small. We chose GPT-2 Small as opposed to other larger pre-trained LMs because GPT-2 Small's surprisal estimates have been shown to best predict human reading time data among the GPT family (Oh and Schuler, 2023), suggesting that it is a more plausible candidate for a model that captures human linguistic knowledge than its relatives. Furthermore, it has been shown that GPT-2 demonstrates more human-like performance in forced-choice judgment tasks with minimal pair sentences involving island violations than other LLMs such as LSTM and Transformer-XL (Warstadt et al., 2020).[3]

2.1 Method and Procedure

We obtained the SLOR values for a wide range of sentences selected from recent studies that reported human acceptability judgment results (examples shown in Table 1). All SLOR values were calculated based on the surprisal estimates for the test sentences from a pre-trained GPT-2 Small model (Radford et al., 2019).[4] If GPT-2 Small indeed captures human linguistic knowledge, and if the SLOR values estimated by LMs constitute a good linking function for sentence acceptability judgments as suggested in previous studies, the computed SLOR values should predict the acceptability judgments from the human experiments.

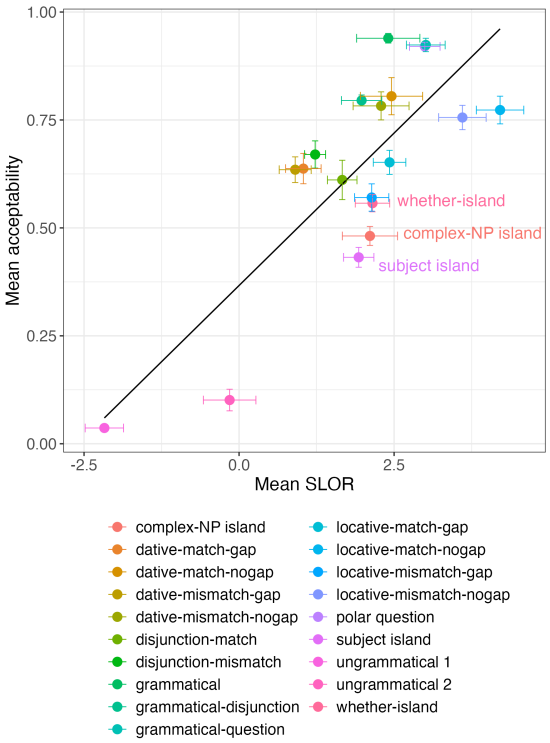


Figure 1: Plot of human acceptability judgments against model SLOR values. Error bars represent 95% bootstrapped confidence intervals. Points representing the three sentence types used as critical conditions in Experiments 2 and 3 (Complex-NP island, subject island, and *whether*-island) are labelled with text.

2.2 Results and Discussion

For the purpose of analysis, all human acceptability judgments from the original studies were linearly

[2]All datasets and scripts can be found here: https://github.com/jmerch/slor-acceptability-satiation.

[3]In Lau et al. (2017), the models tested were 2/3/4-gram models, BHMM, 2-tier BHMM, LDAHMM, and RNNLM, pre-trained on the BNC corpus and the English Wikipedia.

[4]We used the implementation of the 124M-parameter GPT-2 model from the *Transformers* library released by HuggingFace (Wolf et al., 2019).

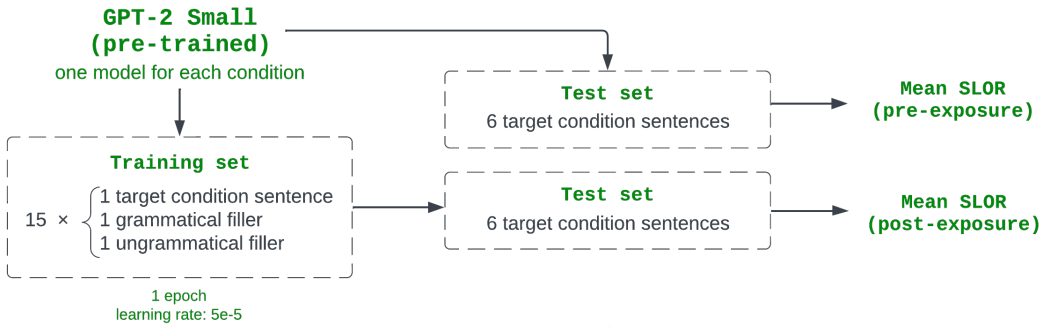


Figure 2: Experimental design of Experiment 2a

Condition	Example	Source
polar question	Does the teacher think that the boy found a box of diamonds?	Lu et al. (2021, 2022)
***whether*-island**	What does the tourist wonder whether the lion attacked __?	
subject island	What does the janitor think a bottle of __ can remove the stain?	
complex-NP island	Who does the king believe the claim that the prince envied __?	
grammatical-question	Who thinks that the doctor decided to treat the mysterious condition?	
ungrammatical 1	Who inspection did not restaurant pass health believes the claim that?	
dative-match-nogap	Kevin gave the children toys and Maria gave the teachers books.	Lu and Kim (2022)
dative-match-gap	Kevin gave the children toys and Maria __ the teachers books.	
dative-mismatch-nogap	Kevin gave the children toys and Maria gave books to the teachers.	
dative-mismatch-gap	Kevin gave the children toys and Maria __ books to the teachers.	
locative-match-nogap	Jacob brushed milk onto the pastry and Emily brushed oil onto the dough.	
locative-match-gap	Jacob brushed milk onto the pastry and Emily __ oil onto the dough.	
locative-mismatch-nogap	Jacob brushed milk onto the pastry and Emily brushed the dough with oil.	
locative-mismatch-gap	Jacob brushed milk onto the pastry and Emily __ the dough with oil.	
disjunction-match	Either Juan or Marie are making the decision.	Lu and Degen (2023)
disjunction-mismatch	Either Juan or these teachers are making the decision.	
grammatical	Julia will perform the surgery tomorrow morning.	
grammatical-disjunction	Either Juan or Marie is making the decision.	
ungrammatical 2	The scientists a discovered solution groundbreaking to	

Table 1: Example stimuli for each sentence type used in Exp. 1. Bolded types are critical conditions used in Exps. 2 and 3.

transformed to a value between 0 and 1 through min-max scaling, with 0 representing the "completely unacceptable" endpoint of the scale, and 1 representing the "completely acceptable" endpoint. Figure 1 shows the mean SLOR values against the mean human acceptability judgments for all the tested sentence types. In a linear regression, SLOR values significantly predicted the human judgments ($\beta = 0.080$, SE = 0.005, t = 17.64, $p < 0.001$), replicating the previous findings reported in Lau et al. (2017). The R^2 value of the model was 0.30, comparable to the best-performing model reported by Lau et al., an RNNLM as implemented by Mikolov (2012), trained on the English Wikipedia, and tested on a set of English Wikipedia sentences after round trip machine translation: $R^2 = 0.32$). The results suggest that the SLOR is a predictor of acceptability. However, we should also note that there is a significant amount of variance in the acceptability data that SLOR failed to capture, challenging the hypothesis that the SLOR values estimated by the pre-trained GPT-2 Small constitute a full linking function for sentence acceptability.

3 Experiment 2a: Deriving Satiation Effects

One crucial property of human acceptability judgments is their malleability: ratings for initially degraded sentences tend to increase with repeated exposure. This effect has been called the "satiation effect" and has been reliably replicated in various sentence acceptability judgment studies (Snyder, 2000; Hiramatsu, 2001; Francom, 2009; Crawford, 2012; Chaves and Dery, 2014, 2019; Brown et al., 2021; Lu et al., 2021, 2022). Crucially, not all sentence types are equally susceptible to satiation: it has been repeatedly observed that certain sentence types resist satiation, and among those that do satiate, satiation rates vary by sentence type (Snyder, 2022; Lu et al., 2023). For example, complex-NP

island sentences show a lower satiation rate than other island sentences, such as subject and *whether*-island sentences (examples shown in Table 1).

In Experiment 2a, we further test whether SLOR provides a good linking function for acceptability judgments in two ways: first, by examining whether SLOR values exhibit the satiation effect (like human acceptability judgments); and second, by investigating whether the varying rates of satiation of different sentence types are predicted by changes in SLOR values after fine-tuning. We follow van Schijndel and Linzen (2018) in using fine-tuning to induce change in surprisal-based metrics from LMs, though our study differs from theirs in that we are interested in the linking function from surprisal to acceptability judgments, rather than to reading times.

3.1 Method and Procedure

This experiment aims to replicate the satiation experiment reported by Lu et al. (2021) using GPT-2 Small. In Lu et al. (2021), human participants were asked to rate the acceptability of three different types of sentences that violated island constraints: complex-NP island sentences, subject island sentences, and *whether*-island sentences. The ratings for all three sentence types increased with increasing presentation order, thus demonstrating the satiation effect. The results from Lu et al. (2021)'s human experiment are shown in Figure 3a.

Importantly, the complex-NP island sentences showed a lower rate of satiation than the other two sentence types. Although it is unclear what makes the complex-NP island sentences satiate at a slower rate, this rate difference has been observed repeatedly and is unlikely to be an artifact of the design (Lu et al., 2022, 2023).

To simulate the repeated exposure in acceptability judgment experiments that gives rise to satiation effects, we fine-tuned GPT-2 Small models using the sentences from Lu et al. (2021). The schematic sketch of the experimental design is shown in Figure 2. For each of the three island types, we fine-tuned a GPT-2 Small model with 45 sentences from the human experiment, consisting of 15 grammatical fillers, 15 ungrammatical fillers, and 15 critical island sentences. The motivation for including the fillers in the training set was to simulate the human experimental experience as closely as possible.

3.2 Results and Discussion

Figure 3b shows the pre- and post-exposure SLOR values. The model-estimated post-exposure SLOR values were higher, by a factor of almost 3, than the pre-exposure values for all three sentence types. This suggests that GPT-2 demonstrates satiation-like behavior in response to exposure to degraded sentences. However, the relative ranking of satiation rates observed in the human results (Figure 3a) was not replicated: in the human experiment, complex-NP island sentences exhibited a significantly lower satiation rate than the other two sentence types; in contrast, the SLOR values for complex-NP sentences increased at a similar rate as *whether*-island sentences, which was higher than the subject island sentences. Thus, the change in SLOR values from by fine-tuning does not reflect the qualitative patterns of change in acceptability ratings through satiation beyond generally showing an increase. This poses a challenge to the proposal to treat SLOR values estimated from LMs as a full linking function for acceptability judgments.

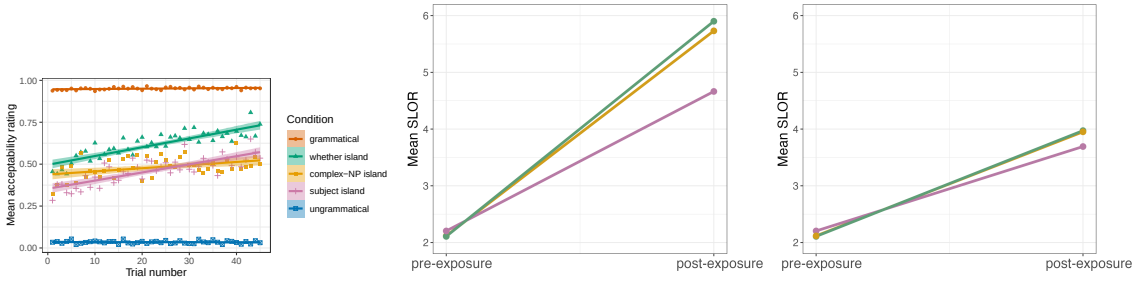
However, there is a caveat to the interpretation of these results: the sentences used for fine-tuning and the sentences in the post-exposure test set contained considerable lexical overlap. In particular, all the complex-NP island sentences from Lu et al. (2021) contained the word sequence "…believe the claim that …". There was much less lexical overlap between training and test sentences in the other two conditions. It is thus possible that the large increase in SLOR for the complex-NP island condition was mostly driven by lexical repetition. To test this hypothesis, we adopt the same design as Experiment 2a in Experiment 2b but with a modified set of training sentences that controlled for lexical repetition.

4 Experiment 2b: Lexical Repetition Control

In this experiment, we test whether the model satiation pattern observed in Experiment 2a persists when we adopt a modified set of training sentences that control for lexical repetition.

4.1 Method and Procedure

The same design as Experiment 2a was adopted. The only difference was that the training set sentences were modified to maximally reduce the repetition of lexical items without changing the sentence's syntactic structure. Whereas the complex-



(a) Plot of mean acceptability by trial number from the human experiment reported in Lu et al. (2021).

(b) Pre- and post-exposure SLOR values in Exp. 2a.

(c) Pre- and post-exposure SLOR values in Exp. 2b.

Figure 3: Comparison of human acceptability judgments reported in Lu et al. (2021) showing satiation effects (a), and model results from Exps. 2a and 2b (b-c).

NP sentences in the original training set all contained the word sequence "believe the claim that", the complex-NP sentences in the modified training set all contained different matrix predicates. Similar modifications were also applied to the subject island sentences and the *whether*-island sentences to minimize lexical repetition (see the design schema in Figure 4).

4.2 Results and Discussion

The pre-exposure and post-exposure SLOR values for all three island sentence types are shown in Figure 3c. The SLOR values increased for all three sentence sentence types post-exposure by about a factor of 2, i.e., at a lower rate than in Experiment 2a. This suggests that lexical repetition did indeed contribute to the large satiation rates observed in Experiment 2a. However, the relative order of the three sentence types' SLOR increases remained the same as in Experiment 2a: the SLOR increase for the complex-NP sentences was comparable to that of the *whether*-island sentences, and higher than that of the subject island sentences. Thus, the comparatively lower satiation rate for complex-NP island sentences observed in the human results was once again not replicated.

In sum, the results from Experiments 2a and 2b demonstrate that GPT-2 Small exhibits satiation-like behavior with repeated exposure to degraded sentences. However, the magnitude and particular patterns of the SLOR increase do not mirror the human satiation effects. There are at least two potential explanations for this discrepancy. First, it is possible that the cognitive processes underlying the satiation effect observed in humans is qualitatively different from the fine-tuning process for LMs. Second, it is possible that the set of linguistic features that affect human satiation are different from the ones that GPT-2's surprisal estimate is sensitive to. Either way, these results challenge both the hypothesis that LM-derived SLOR estimates provide a full linking function for human sentence acceptability judgments, as well as the idea that GPT-2 Small fully captures human linguistic knowledge.

5 Experiment 3: Generalizing Satiation Effects

Another key property of human sentence acceptability judgments is that the acceptability increase gained through satiation generalizes across syntactically related sentence types (Lu et al., 2022). In a series of acceptability judgment experiments employing the same exposure-and-test paradigm as described above, Lu et al. (2022) exposed participants to one of three sentence types: subject island sentences, *whether*-island sentences, and polar questions. In the test phase, participants were asked to rate the acceptability of either subject island sentences or *whether*-island sentences. Exposure and test sentence types were fully crossed. The results are shown in Figures 6a and 6b. Conditions where participants were exposed to one island sentence type and tested on the other (e.g., exposed to subject island sentences and tested on *whether*-island sentences) are labeled "between-category"; conditions where participants were exposed to and tested on the same sentence type are labeled "within-category". Acceptability ratings on test sentences were lower in the between-category than in the within-category condition, but significantly higher than in the control condition, where participants were exposed to polar questions (i.e., non-island sentences) and tested on island sentences. Lu et al. (2022) concluded from these results that the abstract linguistic features shared between the two

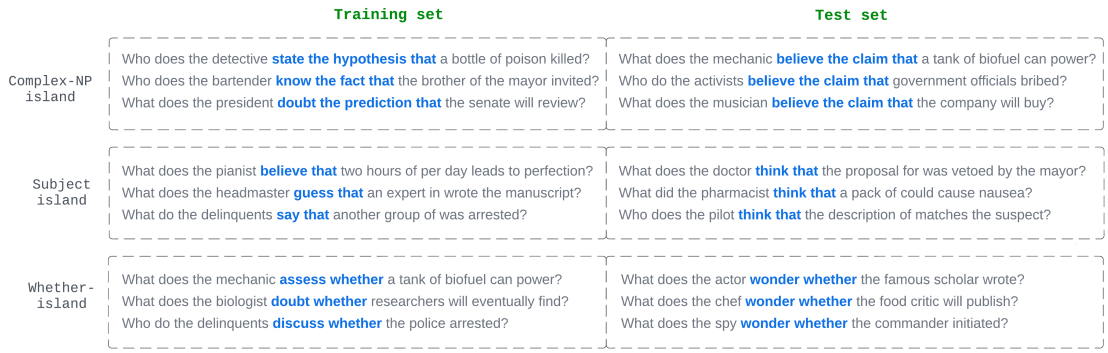


Figure 4: Modified training and test sets used in Experiment 2b to control for lexical repetition

syntactically-related island sentence types (e.g., the existence of long-distance wh-movement, the existence of dependencies violating the subjacency condition, and others) can be used by participants as representational targets for satiation. The polar question sentences are less syntactically similar to the island sentences than the island sentences are to each other. As a result, when participants were exposed to polar questions in the exposure phase, there were fewer shared linguistic representations between the exposure and test sentences that could serve as representational targets for satiation, thus resulting in a smaller satiation generalization effect.

In this experiment, we adopted a similar design as Lu et al. (2022)'s human experiment with GPT-2 Small, with the aim to test whether the SLOR value estimates demonstrate the satiation generalization effect.

5.1 Method and Procedure

The schematic sketch of the experimental design is shown in Figure 5. We fine-tuned a pre-trained GPT-2 Small model with 12 exposure sentences (one of the three sentence types: subject island sentences, *whether*-island sentences, and polar question sentences) and 12 fillers in the training phase. In the test phase, we calculated the fine-tuned models' SLOR estimates for two test sets consisting of subject island and *whether*-island sentences respectively. If the model demonstrates human-like satiation generalization effects, the post-exposure SLOR values should be higher than the pre-exposure values, the SLOR increase in the between-category condition should be smaller or equal to the SLOR increase in the within-category condition, and the SLOR increase in both the between- and within-category condition should be larger than in the control condition.

5.2 Results and Discussion

The results of Experiment 3 are shown in Figures 6c and 6d. In both test sets, the post-exposure SLOR values were higher than the pre-exposure SLOR values (indicated by the dashed lines in the figures) for all conditions. The SLOR increase for the between-category condition is numerically smaller than the within-category condition, similar to the pattern observed in the human results.

However, there was one unexpected observation. The SLOR increase for the control training condition (i.e., when the model was fine-tuned on polar questions and tested on either of the island sentence types) was comparable to the between-category condition when the model was tested on *whether*-island sentences, and even numerically larger than the between-category condition when the model was tested on subject island sentences. This suggests that for the model, the satiation generalization effect from polar questions to the island sentence types was comparable to, if not larger than, the satiation generalization between the two syntactically closely related island sentence types. By contrast, in the human results reported by Lu et al. (2022), the satiation generalization effect from polar questions to island sentences was the smallest among all training conditions.

In sum, we observed satiation generalization effects in the SLOR values estimated by GPT-2 Small. However, the control condition (i.e., the model fine-tuned on polar questions) showed an unexpectedly large satiation generalization effect that was even numerically larger than the between-category condition (at least when testing on subject island sentences). This suggests that the model treats the polar questions as more similar to the subject island sentences than the *whether*-island sentences.

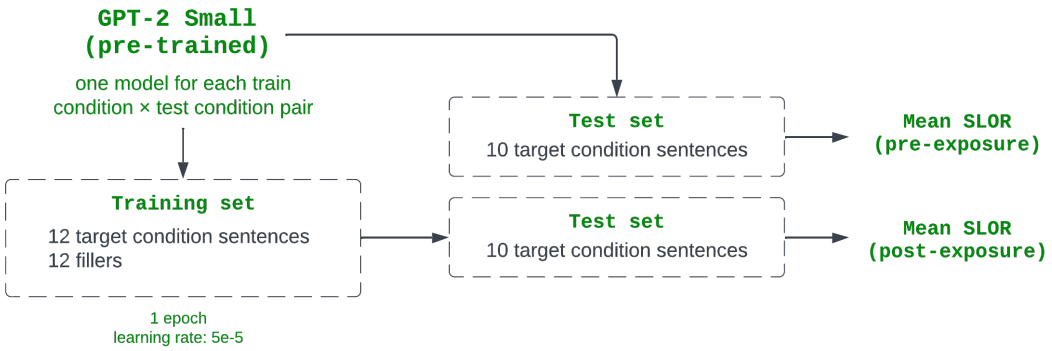


Figure 5: Experimental design of Experiment 3

By contrast, the human results suggest that there are more shared representations between the two island sentence types than between polar questions and either of the island sentence types. There are several possible explanations for this difference between the human and the model results: it is possible that the linguistic features that humans and the model pay attention to during fine-tuning/satiation are different; it is also possible that the three tested sentence types are represented in vastly different ways between humans and the model. Either way, these results again challenge both the hypothesis that LM-derived SLOR estimates provide a full linking function for human sentence acceptability judgments, as well as the idea that LMs fully capture human linguistic knowledge.

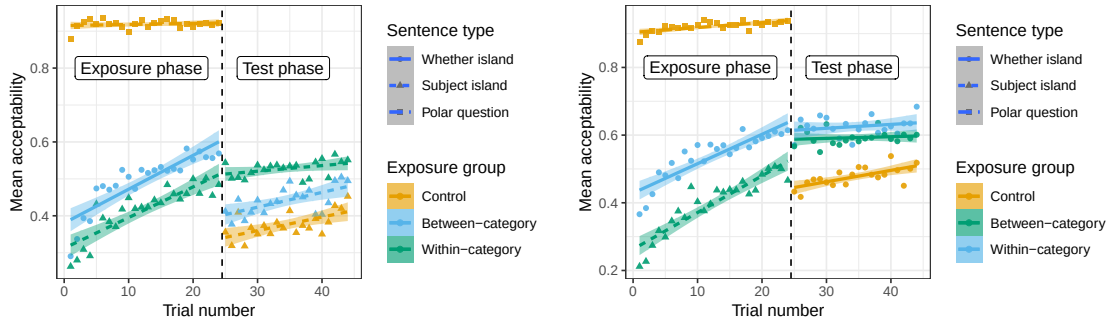
6 General Discussion

In this study, we aimed to test the hypothesis that SLOR values estimated by LMs can provide a linking function for human sentence acceptability judgments. We did so by testing pre-trained GPT-2 Small models in experiments following similar designs as various human sentence acceptability judgment studies, following the recent trend in the computational linguistic literature to treat LMs as subjects in experimental syntax and psycholinguistic experiments (Leong and Linzen, 2023; Futrell et al., 2018, 2019; Wilcox et al., 2023; Arehalli et al., 2022, *inter alia*). We compared the model performance against human results along three dimensions: (1) whether the model-estimated SLOR values predicted human acceptability judgments, (2) whether the increase in SLOR values through model fine-tuning exhibited the same qualitative patterns as the satiation patterns observed in human acceptability judgment experiments exposing participants to degraded sentences, and (3) whether the increase in SLOR values through model fine-tuning exhibited the same qualitative generalization patterns across sentence types as observed in humans.

In Experiment 1, we showed that the SLOR values estimated by the pre-trained GPT-2 Small model predict sentence acceptability judgments given by human participants across a broad range of sentence types, replicating previous results that did not use Transformer models (Lau et al., 2017, 2020). This result suggests that the SLOR values estimated by GPT-2 Small is a plausible linking function for human acceptability judgments broadly. However, there was a lot of variance left unexplained by the SLOR values, suggesting that the linking function proposal is limited.

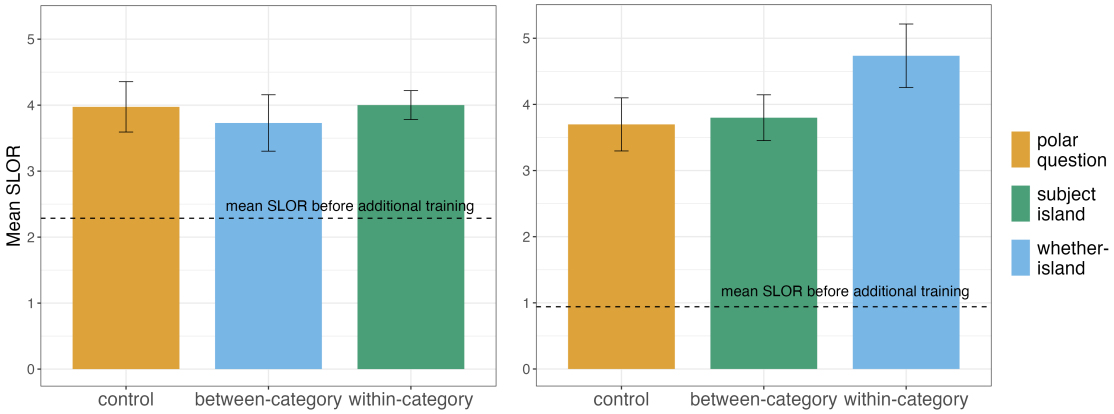
In Experiments 2a and 2b we showed that the SLOR values estimated by GPT-2 Small for degraded sentence types increase when the model is fine-tuned on sentences of the same structure, akin to the satiation effect observed in human participants. However, the magnitude of SLOR increase did not predict the magnitude of acceptability increase for the sentence types we tested. In Experiment 3, we further showed that models fine-tuned on one sentence type showed increased SLOR values for other sentence types, similar to the satiation generalization effect observed in human acceptability judgments experiments. However, the fine patterns of the generalization effect in the models was crucially different from the human results: fine-tuning on polar questions led to a greater SLOR increase for subject island sentences than fine-tuning on *whether*-island sentences, which are more syntactically similar to subject island sentences than polar questions.

In sum, we found that SLOR, a surprisal-based metric, generally predicts sentence acceptability. Fine-tuning LMs as a way of exposing them to



(a) Human results for subject island test condition

(b) Human results for *whether*-island test condition



(c) SLOR values for models trained on various conditions, tested on subject islands

(d) SLOR values for models trained on various conditions, tested on *whether*-islands

Figure 6: Comparison of the satiation generalization effect observed in the human experiments in Lu et al. (2022), shown in sub-figures (a-b), and the model results from Exp. 3, shown in sub-figures (c-d).

novel sentences leads to satiation and generalization effects, but the model results crucially differ from the human results in the fine patterns of the satiation and generalization effects. Our results suggest that LMs, like humans, are sensitive to abstract linguistic representations beyond lexical identity and particular sentence structures. However, the discrepancies with the human results highlight the differences in the relevant linguistic representations or the learning mechanisms between humans and language models, challenging the claim that pretrained LMs like GPT-2 Small can fully capture human linguistic knowledge, or that SLOR estimated by such LMs can fully account for sentence acceptability judgments.

Finally, the results of the current study point to some possible directions for future research. Although we showed that SLOR estimated by GPT-2 does not fully capture human acceptability judgments, this does not definitively reject the hypothesis that surprisal is a causal bottleneck for acceptability (Lau et al., 2017, 2020; Culicover et al., 2022). In order to further investigate the validity of the surprisal bottleneck hypothesis, future studies should examine LMs other than the ones we and the previous literature tested with the aim to gain surprisal estimates that more accurately capture human linguistic knowledge, and also examine metrics other than SLOR that may serve as better linking functions between surprisal and sentence acceptability.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 301–313.

Jessica MM Brown, Gisbert Fanselow, Rebecca Hall, and Reinhold Kliegl. 2021. Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *PLOS One*, 16(5):e0251280.

Rui P Chaves and Jeruen E Dery. 2014. Which subject islands will the acceptability of improve with repeated exposure. In *Proceedings of the 31st West Coast Conference on Formal Linguistics*, pages 96–106.

Rui P Chaves and Jeruen E Dery. 2019. Frequency effects in subject islands. *Journal of Linguistics*, 55(3):475–521.

Jean Crawford. 2012. Using syntactic satiation to investigate subject islands. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 38–45. Cascadilla Proceedings Project Somerville, MA.

Peter W Culicover, Giuseppe Varaschin, and Susanne Winkler. 2022. The radical unacceptability hypothesis: Accounting for unacceptability without universal constraints. *Languages*, 7(2):96.

Elaine Francis. 2022. *Gradient acceptability and linguistic theory*, volume 11. Oxford University Press.

Jerid Cole Francom. 2009. *Experimental Syntax: Exploring the effect of repeated exposure to anomalous syntactic structure–evidence from rating and reading tasks*. Ph.D. thesis, The University of Arizona.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Kazuko Hiramatsu. 2001. *Accessing linguistic competence: Evidence from children's and adults' acceptability judgments.* Ph.D. thesis, University of Connecticut.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science*, 41(5):1202–1241.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

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

Jiayi Lu and Judith Degen. 2023. Acceptability judgment dataset for subject-verb agreement with disjoined subjects. *GitHub repository*.

Jiayi Lu, Michael C Frank, and Judith Degen. 2023. A meta-analysis of syntactic satiation in extraction from islands. *lingbuzz/007198*.

Jiayi Lu and Nayoun Kim. 2022. The puzzle of argument structure mismatch in gapping. *Frontiers in Psychology*.

Jiayi Lu, Daniel Lassiter, and Judith Degen. 2021. Syntactic satiation is driven by speaker-specific adaptation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Jiayi Lu, Nicholas Wright, and Judith Degen. 2022. Satiation effects generalize across island types. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Tomáš Mikolov. 2012. *Statistical language models based on neural networks.* Ph.D. thesis, Brno University of Technology.

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.

Adam Pauls and Dan Klein. 2012. Large-scale syntactic language modeling with treelets. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 959–968.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

John R Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Carson T Schütze. 1996. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press.

William Snyder. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, 31(3):575–582.

William Snyder. 2022. Satiation. In Grant Goodall, editor, *The Cambridge Handbook of Experimental Syntax*, pages 154–180. Cambridge University Press.

Jon Sprouse. 2018. Acceptability judgments and grammaticality, prospects and challenges. In *Syntactic structures after 60 years*, pages 195–224. De Gruyter Mouton.

Martin van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*.