Beyond Surprising: English Event Structure in the Maze

Lisa Levinson*

Abstract. To what extent can we tease apart semantic representations and processes from other influences on processing such as probabilistic prediction? In this paper I detail two experiments testing the hypothesis that there is semantic complexity in the lexical representations of result verbs that influences reaction times *above and beyond* probabilistic distributions. This is done by replicating a self-paced reading study from Levinson & Brennan (2016) while also modelling lexical surprisal. Experiment 1 replicates the original result, but only in experiment 2 using the maze task does the effect emerge beyond surprisals. The more focal maze task results suggest that processing costs associated with bieventive result verbs should be accounted for by grammatical factors, in addition to probabilistic prediction.

1. Introduction. One of the central questions in experimental semantics is how to link the outcomes of various behavioral and neural tasks to the rules and representations of theoretical work. Online tasks such as reaction times and EEG component amplitudes introduce a variety of other influences that must be teased apart from grammatical factors to identify a potentially causal association between a specific grammatical phenomenon and an observed effect. In this paper I explore such questions specifically within the domain of event structure, asking whether there are semantic, event structural properties in the lexical representations of verbs that influence reaction times above and beyond the probabilistic distribution of those verbs and their arguments. More specifically, in this study I test the hypothesis that event structure complexity is associated with a processing cost that cannot fully be explained by lexical surprisal, where surprisal is quantified using the transformer language model GPT-2 (Radford et al. 2019).

Behavioral studies on a variety of event-structure related contrasts in verbs (discussed below) have found processing costs associated with greater complexity. Many such studies have further argued that these costs bolster support for specific semantic event structures. However, many of these studies pre-date important advances in our understanding of the role of prediction in sentence processing and the development of language models which seem to more accurately model human-like predictions. In the spirit of Delogu et al. (2017)'s work on complement coercion, in this paper I revisit the results of Levinson & Brennan (2016)'s experiment 2 on causative event structure in result verbs to test (a) whether the basic findings replicate, and (b) whether the effect of complex event structure goes beyond the effects expected due to sentence prediction, as modelled by surprisals from a transformer language model. This work is part of a series of replications in this vein to improve our understanding of the relationship between event structure, prediction, and sentence processing.

Previous behavioral studies have found "costs" for lexical semantic verb representations due to

*Much gratitude to RAs Yizhi Tang, Lila Tappan, Brighton Pauli, Yasemin Gunal, Emma Thronson, and Thea Kendall-Green who all made valuable contributions to this project! Thanks also to the organizers and participants of ELM 2. This work was supported by University of Michigan's UROP program. Author: Lisa Levinson, University of Michigan (lisalev@umich.edu).

the number of sub-events (McKoon & MacFarland 2000, McKoon & Macfarland 2002, Gennari & Poeppel 2003, McKoon & Love 2011) and event types (Gennari & Poeppel 2003), even in lexical decision where contextual prediction does not play a role. It remains unclear, however, how these effects link with underlying semantic representations, and by what mechanisms they induce such costs.

Structural verb biases (such as frequency of transitive vs. intransitive frames) vary both within and across languages independent of the event structure of the verbs themselves (McKoon & MacFarland 2000, Rappaport Hovav 2020). Event structural properties thus might not travel through the same "causal bottleneck" (Levy 2008) of surprisal, but rather make an independent contribution to processing. The majority of prior findings cannot tease apart these factors; while based on stimuli that are controlled for a variety of probabilistic factors, they have not been recently re-evaluated in the context of (a) probabilities calculated with less sparse language models, (b) measures such as surprisal that are more closely correlated with reading times (Hale 2001, 2016), (c) statistical modeling of multiple stimulus co-variates, and (d) more focal behavioral tasks such as grammatical maze (Freedman & Forster 1985).

2. Background. Levinson & Brennan (2016) tested the hypothesis that even phonologically identical verb forms could show varying processing costs depending on the event complexity associated with their syntactic context. For example, with "result" verbs such as *melt* (called "causative" verbs in the original paper), the same verb form may denote a bieventive causative change-of-state (1), or a monoeventive change-of-state (2).

(1) The sun melted the ice.
(2) The ice melted.

Transitives in this alternation in English have been analyzed in various studies as denoting more complex events than intransitives (inchoatives) (Alexiadou et al. 2006, Pylkkänen 2008, Rappaport Hovav & Levin 2012). Although the details of the number and types of events vary across specific implementations, the hypothesis being tested is whether there is a processing cost due to a greater number of subevents (for example, 2 subevents) in the transitive result verb as compared with it's intransitive variants, which have been more commonly analyzed as denoting a simple event.

In order to control for the potential confound of number of arguments, the study included sentences that also vary in number of arguments, but not number of subevents. These are "manner" verbs (described as "activity" verbs in the original study):

(3) The student read the book.
(4) The student read.

The predicted (and observed) effect of event structure was thus an interaction such that transitive result verbs as in (1) are associated with longer reading times than intransitive result verbs (2) to a greater degree than transitive manner verbs (3) might take longer than intransitive manner

verbs (4).

2.1. ESTIMATING HUMAN PREDICTION. Although it had been well established for many years that the probability of a word given prior context plays an important role in word and sentence processing, Hale (2001) proposed the idea of quantifying this in terms of cognitive load measured as word surprisal. For example, given the partial sentence or prefix "After work, I like to", one can extract from a predictive language model the conditional probability of that string, and then the probability of that string plus another word, such as "sleep". As described in more detail in Hale (2016), these prefix probabilities can then be mathematically transformed into surprisal values. Surprisal inverts the scale so that higher values are less expected, and also log transforms them. These transformed values derived from conditional probabilities have been shown to correlate well with various behavioral and neurolinguistic measures, including self-paced reading (Levy 2008).

Surprisal values can be computed from conditional probabilities derived from a variety of sources, from cloze task (Taylor 1953) measurements to corpus-based bigram and large transformer models. Smith & Levy (2011) demonstrate that cloze probabilities differ significantly from corpus statistics. Michaelov et al. (2021) further show that even when measuring all predictions in surprisal, computational language models more accurately predict the human N400 response, an EEG component strongly associated with expectation and prediction during sentence comprehension. Put together, this work suggests that the best way to currently model the predictions that humans are making during sentence processing is with probabilities extracted from language models and quantified as surprisals.

In the paper being replicated here, several standard measures were taken to attempt to control for non-event-structure influences on sentence processing, both in stimuli creation and the statistical analysis. These included frequency estimates derived from a corpus and transitivity biases estimated by ngram searches of Google Books, and norming to match sentence acceptability across conditions. However these lexical- and sentence-level statistics do not provide very direct insight into the probability or predictability of the verb and following words in the specific sentence context. For example, the verb "knit" is assigned a 66% transitivity probability by Gahl et al. (2004). However, when presented with the sentence context "After work, I like to knit", only 3 of the top 10 GPT-2 predictions are transitive continuations.

2.2. TASKS FOR INVESTIGATING THE ROLE OF PREDICTION. Prior behavioral work on event structure has used tasks such as lexical decision (Gennari & Poeppel 2003), whole sentence reading times (McKoon & MacFarland 2000, McKoon & Macfarland 2002), self-paced reading (Gennari & Poeppel 2003, Brennan & Pylkkänen 2010), and a variant of self-paced reading called the stop-making sense task (McKoon & Love 2011). These methods provide various challenges to teasing apart the role of prediction from other influences on the outcome variables. Single word lexical decision and whole-sentence reading don't provide the context and granularity to explore questions of prediction. Self-paced reading is more granular, but spillover effects make it difficult to separate responses to a current word from those for the prior word. The stop making sense task (Mauner et al. 1995) requires participants to make a deeper decision when proceeding to the next word, as they are asked to specifically detect anomalies - when the sentence stops making sense. This task seems to have potential to address the challenge of spillover and incrementality, depending on the

design of "sense" violations (semantic, syntactic, or lexical) and the number and difficulty of distractor sentences. However, few studies have been conducted using this paradigm thus far. Another task which addresses these challenges in a similar fashion but has a larger literature demonstrating its outcomes and comparison with other tasks is the maze task (Freedman & Forster 1985).

The maze task is detailed in Forster et al. (2009) as an alternative to self-paced reading that leads the participant to more incremental processing. In this task, rather than pressing a button to advance to the next word in the sentence, the participant must first choose the best sentence continuation from two presented choices. The outcome variable is still the response time for the next word in the sentence, but this response now requires an additional discrimination between possibilities. An example of subsequent screens in maze task presentation is given in Figure 1.

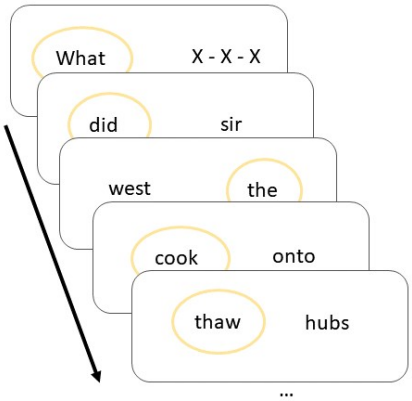

Figure 1: Maze task frame sequence

The distractor alternatives can be designed to be "incorrect" continuations on various dimensions. In the grammatical maze, or g-maze, the alternatives are intended to be ungrammatical continuations. For example, in the context of the English words "That zebra chases", the word "approves" would be an ungrammatical continuation, and might serve as a distractor for the grammatical option of the same length, "leopards". In the lexical maze, or l-maze, the distractors are not known words, so the distractor might be a pseudoword such as "bunklet".

Although the maze task is less naturalistic than self-paced reading, it has several apparent benefits for studying certain phenomena that justify this trade-off. Forster et al. (2009) showed that the task is sensitive to syntactic complexity, and also detail advantages such as lack of spillover effects, and the necessity for participants to commit to a parse at each decision point. This in turn ensures that words are being more fully integrated prior to each button press, rather than delayed to a later point in the sentence. They also found robust effects similar in several respects to eye tracking, though with more incrementality due to the inability to return to earlier portions of the text. Boyce et al. (2020) further demonstrated that the task can be successfully implemented in web-based studies, and that it has greater statistical power and effect "localization" (to a word in the sentence) than even in-lab self-paced reading.

3. Experiment 1. Experiment 1 sought to directly replicate effects of crossing event complexity and transitivity in English (experiment 2 of Levinson & Brennan 2016), using the same method

Table 1: Stimuli for Experiments 1 and 2 (from @levinson.brennan2016a)

exx	w1	w2	Det	N	V	V+1	V+2	w8	Args	Verb Type
1	What	did	the	**cook**	thaw	in	the	cafeteria?	2	Result
2	When	did	the	**popsicle**	thaw	in	the	cafeteria?	1	Result
3	What	did	the	**teacher**	hum	for	the	students?	2	Manner
4	When	did	the	**teacher**	hum	for	the	students?	1	Manner

of self-paced reading, with additional predictors in the statistical analyses to evaluate the relative contribution of event complexity vs. surprisal. The replication was also conducted online rather than in-lab (the setting of the original study).

3.1. Method.

3.1.1. Participants.

90 American English readers completed the study. Participants were undergraduate students at Oakland University and were compensated with course credit.

3.1.2. Materials.

The stimuli were the same as those used in experiment 2 of Levinson & Brennan (2016), where conditions were normed and matched for acceptability but subject animacy varies to allow for this matching. Verb frequency was also matched across VerbType conditions. As seen in Table 1, the conditions cross verb type (result vs. manner) with number of arguments.[1] More details about the stimuli design can be found in the original paper.

There were 87 experimental sentence pairs (43 result, 44 manner). The fillers used were different from those used in the original study, but followed a similar pattern. 36 were ungrammatical questions, to balance out the experimental items for the acceptability task. There were 60 additional declarative sentence fillers with varying ratios of ungrammatical sentences per participant, depending on the stimuli list. For the experimental materials, a Latin square design was used to create two lists such that no participants would see both items in a verb pair.

3.1.3. GPT-2 Surprisals.

Surprisal for each non-initial word in the sentences was estimated using probabilities generated by the open source transformer language model GPT-2 (Radford et al. 2019), as shown for the regions surrounding the verb in Figure 2. These surprisals were calculated in Python using the minicons package (Misra 2022), which provides convenience wrappers for the Hugging Face transformers library (Wolf et al. 2020).

[1] Full stimuli are available via the OSF repository: https://osf.io/zh3ub/.

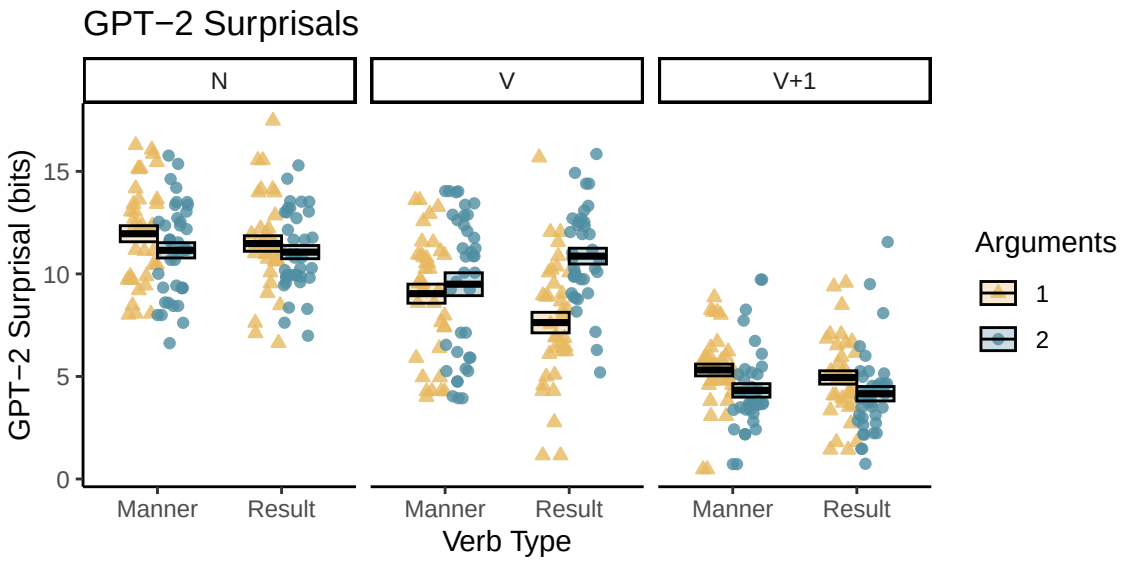

Figure 2: Surprisals for each item by word position from the pre-verbal noun to first word following the verb. Crossbars show mean and standard error of the mean.

3.1.4. PROCEDURE. Participants completed a self-paced moving window task (with post-sentence acceptability judgments) presented online using Ibex presentation software hosted on IbexFarm (Drummond 2013). Participants used the space bar to reveal the sentence word-by-word at their own pace. After the final word in the sentence was presented, a question mark appeared. The subject was instructed to respond press the 'f' key if the sentence was natural, and the 'j' key if the sentence was unnatural. Before the main experiment, participants completed 10 practice trials with feedback for acceptability judgment accuracy.

3.1.5. DATA ANALYSIS. Prior to analysis, trials that were judged to be unacceptable were removed. Participants and items with mean accuracy below 70% on acceptability judgments were also excluded - six participants (out of 90, 7%) and 12 experimental items (out of 87, 14%). No trials were excluded on the basis of reading times, as analyzing log transformed reading times minimized the impact of outliers.

The critical regions analyzed were the verb and the spillover region at verb+1 (preposition). A linear mixed effects models (Gelman & Hill 2006, Baayen et al. 2008) was fit at the verb position, with a model similar to that used in the replicated study, which I will call the "event structure" model (5). This model included fixed effects for the interaction between VerbType and Causativity, using treatment coding for the categorical predictors (with 'manner' as the reference level for VerbType, and 'intransitive' for Transitivity). Also included were fixed effects for centered and scaled lexical statistics and random intercepts for participants and items.[2] Frequency was sourced

[2]Models with random slopes did not converge consistently across different models. Bayesian models with equivalent formulas plus random slopes for VerbType, Transitivity, and surprisal (where relevant) were fit with flat priors in brms (Bürkner 2017) with Stan (Team 2022) to serve as a pseudo-maximum likelihood estimate that is most comparable to those fit by lme4. The coefficients from these models did not substantively differ from those fit by lme4.

from the Corpus of Contemporary American English (COCA) (Davies 2008).

For the V+1 spillover region, where significant effects were observed in the original study, a full model was also fit including a fixed effect of GPT-2 surprisals, as in (6). Finally, a "surprisal" model was fit omitting the event structure interaction (7).

(5) log(rt) ~ VerbType * Transitivity + frequency + length + (1 | participants) + (1| items)
(6) log(rt) ~ VerbType * Transitivity + verb_surprisal + verb_frequency + verb_length + (1 | participants) + (1|items)
(7) log(rt) ~ verb_surprisal + verb_frequency + verb_length + (1 | participants) + (1|items)

Since the effect in the spillover region would be associated with the verb, and not the matched post-verbal words, the surprisals, frequency, and length factored into the analysis of spillover regions were for the verb, not V+1 (a very short preposition). Models were fit using the lme4 package (Bates & Maechler 2009) in R (R Development Core Team 2006), and p-values estimated using the implementation of Satterthwaite approximation provided by the lmerTest package (Kuznetsova et al. 2017).

3.2. RESULTS. The results are visualized word-by-word for the verb and two words following in Figure 3. As in the original study, there was no significant interaction at the verb. Results for the spillover region supported replication of the predicted interaction in the event structure model (β = .03, se = .018, p = .046). A posthoc pairwise comparison showed a significant effect for transitivity in the result verbs as well (β = .03, se = .01, p = .009).

However, when verb surprisal was added to the model (full model), only surprisal evidenced a significant effect (β = .02, se = .008, p = .028). While model comparison between the event structure model and full model showed that GPT-2 significantly improved model fit (likelihood ratio test, p = .025), the addition of event structure in the full model did not improve fit over the surprisal model.

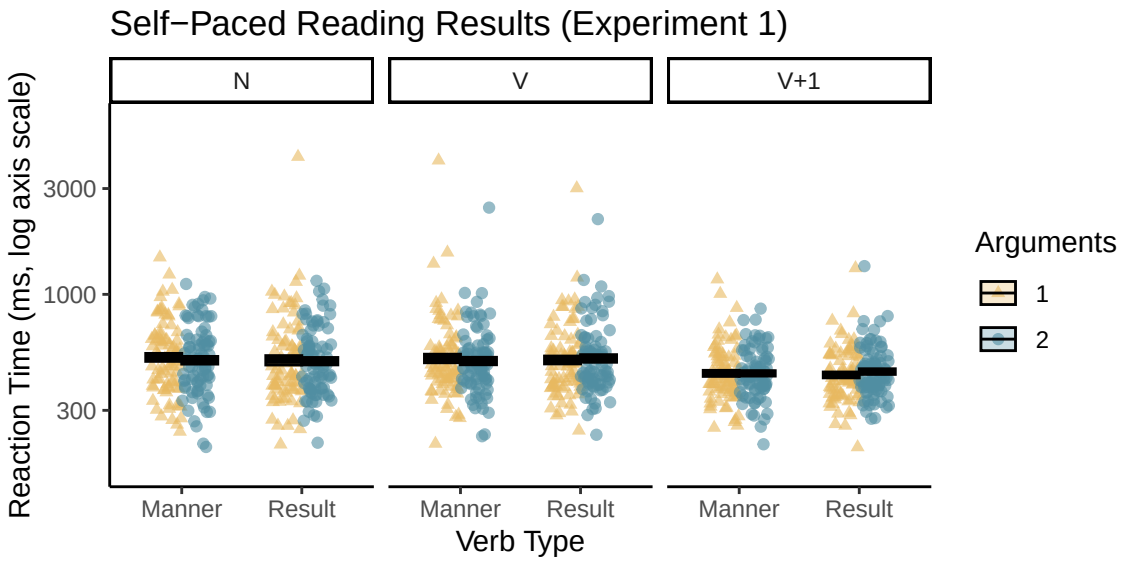

Figure 3: Mean participant reaction times by word position from the pre-verbal noun to first word following the verb. Crossbars show mean and standard error of the mean.

3.3. EXPERIMENT 1 DISCUSSION. The critical interaction effect at the verb+1 region from Levinson & Brennan (2016) did replicate, in that a similar model omitting the predictor of surprisal did provide support for an effect of event structure in transitive result verbs that was not present in the manner verbs. However, once surprisal was added to the full model, it seemed to absorb the effect attributed to event structure.

This does not mean that event structure does not play a role, or that the analysis of result or manner verbs is incorrect. Crucially, surprisal is not a measure that is completely independent from semantic representations. The predictions generated by language models are built from probability distributions that are themselves a product of human language production. Thus, as discussed above, surprisal is a "bottleneck" through which semantic and other grammatical representations are transformed into predictions.

That said, the results of experiment 1 are also consistent with the hypothesis that grammatical contrasts in the stimuli do, to some extent, evade the surprisal bottleneck and more directly influence sentence processing. In self-paced reading, this influence may be obscured by the dispersion of processing costs across multiple words in the spillover region.

4. Experiment 2. Experiment 2 was designed using the maze task in order to encourage participants to read the stimuli sentences more incrementally and carefully, in order to minimize the dispersion of the influence of specific lexical items and maximize the effect at the verb. Concentrating the verb-triggered processing to the verb response time should allow for a more focalized comparison between the role of semantic and other grammatical word properties and probabilistic measures such as surprisal.

4.1. METHOD.

4.1.1. PARTICIPANTS. 60 participants with English as their first language and living in the United States were recruited via Prolific. Payment was $3.33 for approximately 20 minutes to complete the study, for an hourly rate of $10. Ages ranged from 18 to 81, with a mean age of 35.

4.1.2. MATERIALS. The experimental stimuli sentences were the same as those in experiment 1, but word alternatives were also generated for the maze task. These alternatives were generated using the A-maze package (Boyce et al. 2020) with the included pre-trained GRNN language model (Gulordava et al. 2018). After automatic generation, the alternatives were manually checked and adjusted. Where the alternative did not seem sufficiently ungrammatical or inappropriate, it was changed to another alternative with the same word length that was deemed "incorrect" based on experimenter intuition (as all maze alternatives were generated in the original studies using this method). Since the same materials were used for both experiments, stimuli measures such as GPT-2 surprisals and frequencies are also identical to those gathered for Experiment 1.

Since there is no offline acceptability judgment in the maze task, all fillers were grammatical sentences. The experimental stimuli were split into 4 lists (rather than 2) to reduce study completion time, and each participant saw 43 or 44 experimental stimuli plus 40 grammatical declarative sentence fillers.

4.1.3. PROCEDURE. The experiment was run online via IbexFarm using the Ibex stimuli generation scripts from the A-maze package (Boyce et al. 2020).[3] To implement the maze task, each sentence is presented in a series of frames. Each frame shows two sentence continuation candidates for the participate to choose from, to the left and right of center. The correct continuation is randomly assigned to a side and varies for individual items across participants to counterbalance. Since there is no "continuation" at the first word, the first alternative is simply "x-x-x". If a participant made the correct selection, the next frame would be displayed. If they chose the alternative, they were presented with error feedback and the rest of that item was not presented. The time from the presentation of the frame until their selection is recorded and serves as the primary outcome variable, the response time. The session started with instructions followed by 3 practice trials with positive and negative feedback. For the experimental materials, a Latin square design was used to create lists where no participants would see both items in a verb pair. Participants were given a 30-second break every 12 sentences.

4.1.4. DATA ANALYSIS. All trials with incorrect maze responses were excluded from the analysis. Since trials ended whenever an incorrect choice was made, subsequent words in the same sentence were not presented.

The same range of event structure, surprisal, and full models were fit as for the self-paced reading results, but only at the verb since there was no evidence of spillover effects across the maze experiment. For these data, it was also possible to add a random slope of surprisals for participants (in models including surprisal) and transitivity for participants, so the full model was as in (8).

[3] Although the scripts use the original Ibex controllers and the maze controller developed by Boyce et al. (2020), a clonable demo of the experiment is hosted on the PCIbexFarm (Schwarz & Zehr 2021) and linked from the OSF repository: https://osf.io/zh3ub/.

(8) log(rt) ~ VerbType * Transitivity + verb_surprisal + verb_frequency + verb_length + (1 + verb_surprisal | participants) + (1 + Transitivity |items)

4.2. RESULTS. As predicted, the maze results exhibited more focal effects, with no apparent spillover, as can be seen in Figure 4. Even with the full model (including surprisal along with event structure predictors), the interaction (β = .13, se = .04, p = .002) was significant at the verb. The predicted pairwise effect for transitivity in causative verbs was not significant in the full model (β = .08, se = .05, p = .07), but a likelihood ratio test comparing that model to one omitting transitivity suggested significantly improved fit from the event structure predictors (p < .001). Model comparison between the full model and partial models showed that both event structure predictors (LRT p < .001) and surprisals (LRT p < .001) improved model fit over either event structure or surprisals "alone".

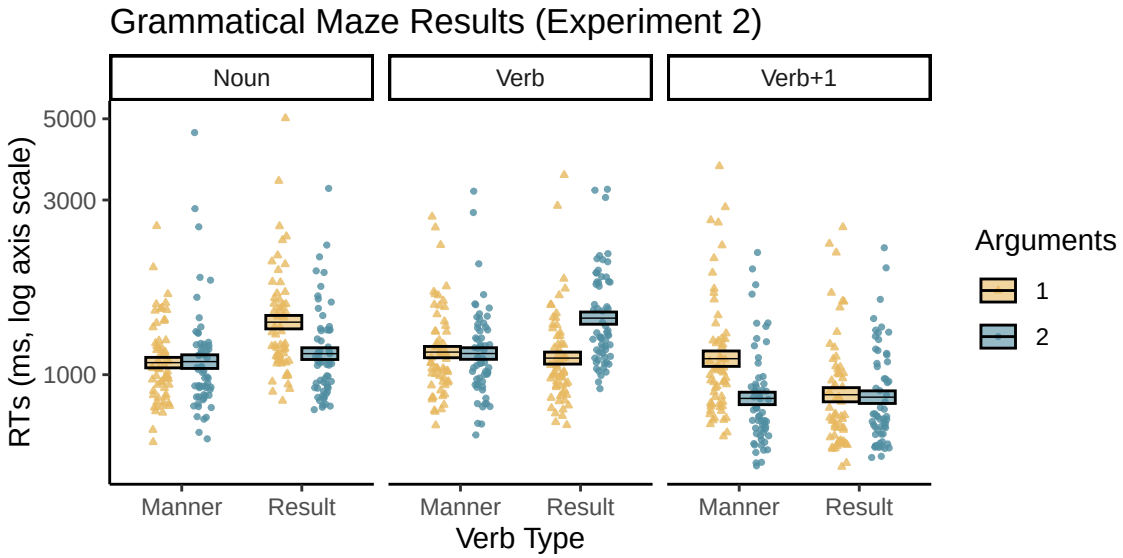

Figure 4: Mean participant reaction times by word position from the pre-verbal noun to first word following the verb. Crossbars show mean and standard error of the mean.

4.3. EXPERIMENT 2 DISCUSSION. Effects of event complexity beyond surprisal are evident in the maze task data. This not only supports the hypothesis that some event structural properties evade the surprisal bottleneck, but also demonstrates that a more incremental task such as maze can help to tease apart these variables.

5. Conclusion. In conclusion, these results support an independent contribution of event structure complexity to incremental processing above and beyond surprisal in the slower but more incremental maze task. Comparison of methods suggests that such effects may only be separable with more focal and larger effects that allow for teasing apart multiple fine-grained contributions to sentence processing.

References.

Alexiadou, Artemis, Elena Anagnostopoulou & Florian Schäfer. 2006. The properties of anti-

causatives crosslinguistically. In Henk van Riemsdijk, Harry van der Hulst, Jan Koster & Mara Frascarelli (eds.), *Phases of Interpretation*, vol. 91, 187–212. Berlin, New York: Mouton de Gruyter. doi:10.1515/9783110197723.4.187.

Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4). 390–412. doi:10.1016/j.jml.2007.12.005.

Bates, Douglas & Martin Maechler. 2009. Lme4: Linear mixed-effects models using S4 classes.

Boyce, Veronica, Richard Futrell & Roger P. Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language* 111. 104082. doi:10.1016/j.jml.2019.104082.

Brennan, Jonathan & Liina Pylkkänen. 2010. Processing psych verbs: Behavioral and MEG measures of two different types of semantic complexity. *Language and Cognitive Processes* 25(6). 777–807. doi:10.1080/01690961003616840.

Bürkner, Paul-Christian. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80. 1–28. doi:10.18637/jss.v080.i01.

Davies, Mark. 2008. Word frequency data from The Corpus of Contemporary American English (COCA). https://www.wordfrequency.info.

Delogu, Francesca, Matthew W. Crocker & Heiner Drenhaus. 2017. Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition* 161. 46–59. doi:10.1016/j.cognition.2016.12.017.

Drummond, Alex. 2013. Ibex farm.

Forster, Kenneth I., Christine Guerrera & Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods* 41(1). 163–171. doi:10.3758/BRM.41.1.163.

Freedman, Sandra E. & Kenneth I. Forster. 1985. The psychological status of overgenerated sentences. *Cognition* 19(2). 101–131. doi:10.1016/0010-0277(85)90015-0.

Gahl, Susanne, Dan Jurafsky & Douglas Roland. 2004. Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments, & Computers* 36(3). 432–443. doi:10.3758/BF03195591.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gennari, Silvia & David Poeppel. 2003. Processing correlates of lexical semantic complexity. *Cognition* 89(1). B27–B41. doi:10.1016/S0010-0277(03)00069-6.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv:1803.11138 [cs]* .

Hale, John. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, doi:10.3115/1073336.1073357.

Hale, John. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass* 10(9). 397–412. doi:10.1111/lnc3.12196.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82. 1–26. doi:10.18637/jss.

v082.i13.
Levinson, Lisa & Jonathan Brennan. 2016. The costs of zero-derived causativity in English: Evidence from reading times and MEG. In Daniel Siddiqi & Heidi Harley (eds.), *Morphological Metatheory*, vol. 229 Linguistik Aktuell/Linguistics Today, 163–198. John Benjamins.
Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. doi:10.1016/j.cognition.2007.05.006.
Mauner, Gail, Michael K. Tanenhaus & Gregory N. Carlson. 1995. Implicit arguments in sentence processing. *Journal of Memory and Language* 34(3). 357–382. doi:10.1006/jmla.1995.1016.
McKoon, Gail & Jessica Love. 2011. Verbs in the lexicon: Why is hitting easier than breaking? *Language and Cognition* 3. 313–330. doi:10.1515/LANGCOG.2011.011.
McKoon, Gail & Talke MacFarland. 2000. Externally and internally caused change of state verbs. *Language* 76(4). 833–858. doi:10.2307/417201.
McKoon, Gail & Talke Macfarland. 2002. Event templates in the lexical representations of verbs. *Cognitive Psychology* 44. doi:10.1016/S0010-0285(02)00004-X.
Michaelov, James A., Seana Coulson & Benjamin K. Bergen. 2021. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *arXiv:2109.01226 [cs, math]* .
Misra, Kanishka. 2022. Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv:2203.13112 [cs]* .
Pylkkänen, Liina. 2008. *Introducing arguments*. Cambridge, MA: MIT Press.
R Development Core Team. 2006. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8). 9.
Rappaport Hovav, Malka. 2020. Deconstructing internal causation. In Elitzur A. Bar-Asher Siegal & Nora Boneh (eds.), *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop* Jerusalem Studies in Philosophy and History of Science, 219–255. Cham: Springer International Publishing.
Rappaport Hovav, Malka & Beth Levin. 2012. Lexicon uniformity and the causative alternation. In *The Theta System*, Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199602513.003.0006.
Schwarz, Florian & Jeremy Zehr. 2021. Tutorial: Introduction to PCIbex – an open-science platform for online experiments: Design, data-collection and code-sharing. *Proceedings of the Annual Meeting of the Cognitive Science Society* 43(43).
Smith, Nathaniel & Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 33, .
Taylor, Wilson L. 1953. "Cloze Procedure": A new tool for measuring readability. *Journalism Quarterly* 30(4). 415–433. doi:10.1177/107769905303000401.
Team, Stan Development. 2022. Stan modeling language users guide and reference manual, version 2.3.
Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi,

Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.