Deconstructing sentence disambiguation by joint latent modeling of reading paradigms: LLM surprisal is not enough

Dario Paape (paape@uni-potsdam.de)[1], Tal Linzen[2] & Shravan Vasishth[1]
[1]Department of Linguistics, University of Potsdam
[2]Center for Data Science/Department of Linguistics, New York University

Abstract

Using temporarily ambiguous garden-path sentences (*While the team trained the striker wondered* . . .) as a test case, we present a latent-process mixture model of human reading behavior across four different reading paradigms (eye tracking, uni- and bidirectional self-paced reading, Maze). The model distinguishes between garden-path probability, garden-path cost, and reanalysis cost, and yields more realistic processing cost estimates by taking into account trials with inattentive reading. We show that the model is able to reproduce empirical patterns with regard to rereading behavior, comprehension question responses, and grammaticality judgments. Cross-validation reveals that the mixture model also has better predictive fit to human reading patterns and end-of-trial task data than a mixture-free model based on GPT-2-derived surprisal values. We discuss implications for future work.

Keywords: garden path; mixture model; multinomial processing tree; surprisal

Background

A range of experimental methods exists to study human reading, eye tracking being the most well-known. Another widely-used method is self-paced reading (SPR), in which the stimulus text is revealed incrementally via key presses on a computer keyboard and the time between key presses is recorded (Just et al., 1982). Classic self-paced reading is unidirectional, that is, rereading is not possible; however, Paape and Vasishth (2022c) recently developed a variant of the paradigm that does allow participants to return to earlier words, called bidirectional self-paced reading (BSPR). Another relatively recent invention is the Maze paradigm (Forster et al., 2009), which requires the participant to "build up" the sentence through sequential binary choices (via keyboard presses) between a grammatical continuation and an ungrammatical alternative (or non-word). Like classic SPR, Maze does not allow rereading.

A crucial tacit assumption made in the field is that all reading paradigms tap into the same set of cognitive processes, modulo differences related to, e.g., motor and eye movement control, parafoveal preview, and the (im)possibility of rereading. Indeed, reading measures collected with different methods tend to correlate, and often (but not always) show converging results when used with the same sets of stimuli (e.g., Frank et al., 2013; Koornneef and Van Berkum, 2006; Paape and Vasishth, 2022a; Witzel et al., 2012). However, to our knowledge, there has been no attempt to build a computational model that captures reading across the different paradigms. We present such a model below, which takes the form of a multinomial processing tree (MPT; see Erdfelder et al., 2009 for a review). In addition to providing a unified account of latent cognitive processes across reading paradigms, our model is probabilistic in the sense that not all processes are assumed to occur in all reading trials. Finally, our model assumes a direct connection between implicit reading measures and behavioral outcomes, that is, acceptability judgments and responses to comprehension questions, thus acknowledging the link between these two types of data at the trial level.

Our approach builds on previous modeling work by Paape and Vasishth (2022b, 2025). Their work focuses on so-called garden-path effects in sentences with temporary syntactic ambiguities. Garden-path sentences are a fruitful target for modeling due to the large variety of reading behaviors that they generate , and due to the range of processing theories that have been proposed (Frazier & Rayner, 1982). An example of a garden-path sentence is given in (1).

(1) While the team trained(,) the striker wondered whether the damage would heal.

In the version of (1) without a comma, readers tend to initially assume that the striker was trained by the team. However, at *wondered*, this analysis becomes untenable, as *the striker* must now serve as the subject of the main clause, and *trained* must be intransitive (as in the comma version). There is converging experimental evidence from different reading paradigms that the no-comma version of (1) is more difficult to read than the comma version, as indexed by longer reading times at the critical word *wondered* and/or, depending on whether the experimental paradigm allows rereading, by more regressions and more rereading of the sentence. The no-comma version is also more prone to misinterpretation, in the sense that participants often answer *yes* to the question *Did the team train the striker?*, and is more likely to be rejected as ungrammatical (see Paape and Vasishth, 2022c for a review).

The standard analytical approach in the field is to statistically compare the mean reading times, regression probabilities, and the proportion of positive acceptability judgments/correct question responses of the comma and no-comma versions across many participants and many experimental items that follow the abstract structure of (1) (e.g, *While the mother dressed the baby cried* . . .). The increase in mean reading time/regression probability in the no-comma compared to

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

the comma version is taken to index the cognitive cost of garden-pathing. Many authors interpret this cost to reflect syntactic reanalysis, that is, the cost associated with abandoning the incorrect syntactic structure and assembling the correct one, or as a reranking of syntactic analyses computed in parallel (e.g., Meng and Bader, 2000). Increases in rejection and/or misinterpretation probability are usually interpreted as processing failures (e.g., Warner and Glass, 1987).

There are two crucial drawbacks to the standard analysis approach. First, analyzing reading times and rejection/misinterpretation probabilities in isolation ignores the assumption that both kinds of measures arise from the same set of cognitive processes (Y. Huang & Ferreira, 2021). Second, by only comparing mean (that is, grand average) reading times between conditions, researchers assume that there is one common distribution that generates all observations. This assumption glosses over the possibility that completely different things may happen in different experimental trials, that is, that garden-path processing may be fundamentally non-deterministic. In some trials, readers may not pay full attention to the stimulus, quickly "skim" the sentence, and then resort to guessing when asked a comprehension question. In other trials, they may notice the incongruity at *wondered* and show a slight reading slowdown, but postpone reanalysis and keep going forward, or not attempt reanalysis at all. Alternatively, they may make a full-scale reanalysis attempt, either via overt rereading or in a "covert" fashion, that is, by stopping stopping at the critical word and mentally backtracking (Lewis, 1998). Finally, even though there is a strong tendency to misanalyze garden-path sentences of the type shown in (1), there may be trials in which the correct analysis is nevertheless initially adopted even in the absence of a comma, so that no reanalysis is necessary. A plausible computational model should incorporate all of these possibilities.

The MPT model

Adopting a non-deterministic view of garden-pathing implies that reading times should follow a mixture distribution. Our model assumes the following five mixture components:

① If not reading attentively:
$(RT - shift) \sim LogNormal(\mu,\ \sigma_1)$

② If reading attentively but not garden-pathed:
$(RT - shift) \sim LogNormal(\mu + att\_cost,\ \sigma_2)$

③ If garden-pathed, no in-situ (covert) reanalysis:
$(RT - shift) \sim LogNormal(\mu + att\_cost + gp\_cost,\ \sigma_2)$

④ If garden-pathed, in-situ (covert) reanalysis:
$(RT - shift) \sim LogNormal(\mu + att\_cost + gp\_cost + reanalysis\_cost,\ \sigma_2)$

⑤ If garden-pathed, regression triggered:
$(RT - shift) \sim LogNormal(\mu + att\_cost + gp\_cost + regression\_cost,\ \sigma_2)$

All cost parameters are restricted a priori to be greater than zero, and σ_1 is restricted to be smaller than σ_2, assuming that reading times registered during inattentive reading are more uniform compared to attentive reading. We model reading times at the critical region as well as the spillover region — that is, the region directly following the critical region — to account for postponed reanalysis (see below). The $shift$ parameter encodes the assumption that part of the total measured reaction time is non-decision time, that is, time required to encode the stimulus, execute motor actions, and so forth (e.g., Rouder, 2005). The model assumes SPR, BSPR and Maze to have the same amount of positive shift, while eye tracking has a smaller shift. Reading times at the critical and spillover region share the same shift, and the amount of shift is assumed to vary between participants (Nicenboim & Vasishth, 2018).

The second component of the model is informed by the presence or absence of rereading in the trial (after having read the critical region), as well as by the acceptability judgment (*accept/reject*) or comprehension question response (*Did the team train the striker?*; *yes/no*) registered at the end of the trial. Accepting the sentence as grammatical or correctly answering the comprehension question with *no* — depending on the task used in the experiment — are assumed to reflect successful parsing of the critical dependency, while rejecting the sentence as ungrammatical or incorrectly answering *yes* are assumed to reflect unsuccessful parsing (but see below for an important difference). Trials are grouped into four categories according to the possible combinations of these two binary variables (*accept/no, no regression; accept/no, regression; reject/yes, no regression; reject/yes, regression*). The four outcomes are modeled with a multinomial distribution whose event probabilities are determined by the structure of the processing tree. The event probabilities are also connected to the mixture distribution of reading times via the mixing proportions, which allows us to estimate process probabilities and process costs in parallel (see also Heck et al., 2018). The model is hierarchical with crossed random effects, that is, all costs and probabilities can vary between participants and sentences.

The multinomial processing tree assumes the following latent processes with their associated probabilities:

1. Attention ($p\_attentive$): The probability that the participant is reading attentively. Varies between participants.
2. Garden-pathing (p_gp): The probability that the participant is garden-pathed. Varies between conditions (comma vs. no-comma), participants, and sentences.
3. Overt reanalysis (p_overt): The probability that the participant engages in overt reanalysis (rereading). If there is no overt reanalysis, it is assumed that covert reanalysis is carried out instead. Varies between participants.
4. Postpone reanalysis ($p\_postpone$): If reanalysis is covert, it can be postponed from the critical to the spillover region, so that the slowdown will register there. Varies between participants.
5. Reanalysis success ($p\_success\_o$/$p\_success\_c$): The probability that overt/covert reanalysis is successful. Varies between garden-path types (see below).
6. Pragmatic inference (p_infer): The probability that the participant engages in pragmatic inference when answering

a comprehension question. For instance, when being asked *Did the team train the striker?* about (1), the participant may plausibly infer that the team did train the striker, even though the sentence did not explicitly say so (e.g., Christianson et al., 2001). Varies between participants and sentences.

7. Baseline regression probability ($p\_base\_regress$): Participants may reread the sentence even in the absence of garden-pathing in order to correct for motor errors or to confirm their interpretation. Varies between participants.

The probability of overt reanalysis and the base regression probability are assumed to be higher in eye tracking compared to BSPR, and p_overt and $p\_base\_regress$ are set to zero for SPR and Maze experiments. The regression cost is assumed to be higher in BSPR compared to eye tracking.

Accounting for pragmatic inferences through p_infer is a novel addition to our MPT model compared to the model of Paape and Vasishth (2025). p_infer is intended to account for an inflated amount of *yes* responses to comprehension questions that are not informative about the actual parsing outcome. The addition of this parameter allows us to simultaneously model data from studies with acceptability judgments and studies with comprehension questions by assuming that (correct) *accept* judgments and (correct) *no* responses map onto each other *modulo* the contribution of p_infer to the question responses.

Our model also contains two simplifications compared to the models of Paape and Vasishth (2022b, 2025). The first simplification is that there is no response bias for the guessing process, that is, guesses are assumed to generate *accept/reject* or *yes/no* responses with equal probability. The second simplification is the removal of the "triage" process assumed by Paape and Vasishth (2022b), which generates quick rejections without reanalysis (e.g., Fodor and Inoue, 2000). Model comparisons by Paape and Vasishth (2025) showed that assuming triage did not improve model fit over a model without triage.

To illustrate how the event probabilities of the multinomial distribution arise from the probabilities of the latent processes assumed by the MPT, we use the probability of observing a trial with a *yes* response — that is, an incorrectly answered comprehension question — and no rereading.

From not observing a regression, we can already conclude that no baseline regression was triggered, and that no overt reanalysis took place. There are four remaining paths through the tree that can generate the observed outcome:

1. Inattentive reading + guessing the response
2. Attentive reading + no garden-pathing + pragmatic inference
3. Attentive reading + garden-pathing + failed covert reanalysis
4. Attentive reading + garden-pathing + successful covert reanalysis + pragmatic inference

The two paths that include covert reanalysis are further split into whether reanalysis was carried out at the critical region or postponed and carried out at the spillover region. The information of whether covert reanalysis was carried out or not, and in which region, comes from the reading-time data via the estimated cost parameters and mixing proportions.

The overall formula for the event probability $p\_(yes + no\_regress)$ looks as follows:

$$
\begin{aligned}
&p\_(yes + no\_regress) = (1 - p\_base\_regress) * \\
&((1 - p\_attentive) * 0.5 + \\
&\quad p\_attentive * (1 - p\_gp) * p\_infer + \\
&\quad p\_attentive * p\_gp * (1 - p\_overt) * (1 - p\_postpone) * (1 - p\_success\_c) + \\
&\quad p\_attentive * p\_gp * (1 - p\_overt) * p\_postpone * (1 - p\_success\_c) + \\
&\quad p\_attentive * p\_gp * (1 - p\_overt) * (1 - p\_postpone) * p\_success\_c * p\_infer + \\
&\quad p\_attentive * p\_gp * (1 - p\_overt) * p\_postpone * p\_success\_c * p\_infer)
\end{aligned}
$$

For grammaticality judgments, the p_infer paths are blocked, and there are thus fewer paths that lead to an (incorrect) *reject* judgment in the absence of rereading: only inattention followed by guessing and unsuccessful covert reanalysis can generate this outcome.

Besides presenting an extended version of the Paape and Vasishth (2022b, 2025) MPT model, we will also compare it against a model based on large language model (LLM) surprisal across different experimental paradigms, and additionally consider a hybrid model in which LLM surprisal affects reading times *in addition* to the mechanisms assumed by the mixture model.

Surprisal as a simpler competitor model

Despite the complexity of human reading behavior, a model is only as good as its fit to the data, and simpler models should be preferred over more complex models if they offer comparable fit. One such model that we will compare against our multi-process model is based on surprisal, that is, the (negative log) probability of a word given the preceding linguistic context (Hale, 2001; Levy, 2008). The basic assumption is that the lower a word's probability, the greater the processing difficulty. Levy (2008) has argued that surprisal is a "causal bottleneck" in the sense that all cognitive processes that create, maintain and alter linguistic representations can only affect "behavorial observables" such as reading times indirectly by changing conditional next-word probabilities.

Under this view, garden-path sentences like (1) are not inherently special: the word *wondered* is simply less predictable when the comma is absent compared to when it is present. Furthermore, given that "all types of hierarchical predictive information present in language — syntactic, semantic, pragmatic, and so forth — must inevitably bottom out in predictions about what specific word will occur in a given context" (Smith & Levy, 2013, p. 313), it should theoretically be sufficient to model to human reading times with a linear regression model with word-level surprisal as the sole predictor, irrespective of the type of structure being studied. With the advent of LLMs, obtaining surprisal values for arbitrary sentences has become relatively easy compared to past studies that used sentence completion tasks, and LLM-derived surprisal often outperforms "empirical" surprisal in terms of fit to reading measures (e.g., Hofmann et al., 2022; Lopes Rego et al., 2024).

Surprisal has, overall, been highly successful at predicting human sentence processing behavior across a variety of experimental methods (e.g., Shain et al., 2024; see Staub, 2025 for a review). However, garden-path sentences such as

(1) have proven challenging for surprisal theory: assuming a linear relationship between surprisal and reading times tends to underpredict the empirically observed difference between the ambiguous and unambiguous versions of the sentence, implying that garden-path sentences *are* in some way special (K.-J. Huang et al., 2024; Van Schijndel & Linzen, 2021).

Paape and Vasishth (2025) used bidirectional self-paced reading data from garden-path sentences to compare the predictive fit of their MPT mixture model to that of several mixture-free models based on surprisal values computed from different LLMs. Using a cross-validation procedure to evaluate predictive fit, they found that the MPT model outperformed the surprisal-based models for their data. Paape and Vasishth took their results to imply that humans recruit mechanisms beyond next-word prediction, such as a specialized syntactic reanalysis mechanism, to process garden-path sentences. If this conclusion proves to be warranted even across reading paradigms, it would be in line with psycholinguistic theories posited decades ago (e.g., Frazier, 1979; Marcus, 1980; Pritchett, 1992).

Fitting the models to reading data from different paradigms

To evaluate our model, we collected a convenience sample of open-access reading data on garden-path processing collected with different reading paradigms. Besides the so-called NP/zero complement (NP/Z) garden path illustrated in (1), we also included data on main verb/reduced relative (MV/RR) garden paths, as illustrated in (2).

(2) a. The girl (who was) fed the lamb remained relatively calm.

b. The lawyer/package sent by the governor was neglected by the officials.

Here, the initial misanalysis is to interpret the first noun (*girl*) as the agent of the first verb (*fed*). At the actual main verb (*remained*), it becomes clear that the noun is, in fact, the patient (*the girl who was fed. . .*). Unambiguous control conditions are created by either making the relative clause explicit as in (2a), or by using an inanimate noun as in (2b).

In order to get crisp estimates of the latent process probabilities and costs, the MPT model optimally needs reading times as well as grammaticality judgments or comprehension question responses in each experimental trial. Our search of the literature yielded relatively few studies that meet these criteria. For instance, we could not find a single Maze study with comprehension questions, as it is typically argued that making the correct choice in Maze presupposes accurate comprehension (Forster et al., 2009). Whether this is true or not is not a question we aim to answer here. In principle, one could formulate the MPT in such a way that comprehension in Maze is assumed to be always accurate; we decided to treat the Maze comprehension accuracy as missing data.

We selected the following studies for our sample:

- K.-J. Huang et al. (2024): A large-scale SPR study on NP/Z and MV/RR garden paths (n = 2000) known as the syntactic ambiguity processing (SAP) benchmark. Comprehension questions about the critical dependency in a subset of trials.
- Y. Huang and Ferreira (2021), Expt. 1: An eye-tracking study on NP/Z garden paths (n = 120). Comprehension questions about the critical dependency in each trial.[1]
- Fujita (2025): A Maze experiment on NP/Z garden paths (n = 60). No end-of-sentence task.
- Paape and Vasishth (2022a): Three experiments (1 × SPR, 2 × BSPR) on NP/Z and MV/RR garden paths (each n = 100). Grammaticality judgments (*accept/reject*) in each trial.

For the eye-tracking and BSPR data, first-pass reading time — that is, the time registered from first entering the region of interest to first leaving it — was used as the measure of interest. The data from all six experiments were preprocessed and combined into one data set in R (R Core Team, 2020). Trials with reading times smaller than 150 ms or greater than 5000 ms at either the critical or spillover region were removed from the data. All models were fitted in Stan (Stan Development Team, 2024) via the CmdStanR interface (Gabry et al., 2024). Informative but not overly restrictive priors based on previous work were used. Two chains with 2000 iterations were run, with the first 1000 iterations being discarded as burn-in. $\hat{r}$ values were monitored for possible indications of non-convergence.

For the surprisal-based regression model, we computed surprisal values for the critical and spillover regions of each sentence using the 124M variant of GPT-2 (Radford et al., 2019) from HuggingFace (Wolf et al., 2019). The surprisal slope can vary between participants. For both the MPT and surprisal models, we also entered region length in characters as a predictor, as the studies used regions of different lengths. LKJ priors (Lewandowski et al., 2009) with $\eta = 4$ were used for the variance-correlation matrices of the random effects.

We also implemented two further models for comparison, bringing the number of models under consideration to four:

1. Baseline: A linear regression model with only region length as a predictor, plus random effects.
2. MPT: The MPT model as described above.
3. Surprisal: A linear regression model with GPT-2 surprisal as a predictor, plus region length and random effects.
4. MPT-plus-surprisal: The MPT model as described above, plus a linear surprisal slope for reading times.

All models assume a lognormal likelihood for the reading time distributions. For the surprisal model (and the baseline), regression probabilities and grammaticality judgments/question responses are treated as Bernoulli variables, and separate surprisal slopes are estimated for all measures.

[1] Both Y. Huang and Ferreira (2021) and Fujita (2025) used an additional manipulation concerning gender mismatches between reflexive pronouns and their antecedents, but the critical region with regard to this manipulation appeared after the regions we analyze here. Differences in earlier sentence regions are reflected in the surprisal values, however.

Note that our MPT model does not include a main effect of experiment or any interactions with experiment. Information regarding which study a given data point comes from only enters into the model indirectly. Participants and sentences are both assumed to be sampled from the same populations across all experiments. We hard-code the experimental method, which determines the amount of shift (non-decision time), the (im)possibility of regressions, baseline regression probability, and regression cost. We also hard-code the end-of-trial task, which determines whether pragmatic inferences are assumed to occur or not. Finally, within the MV/RR experiments, we hard-code the disambiguation type — as shown in (2) above — as it may influence the strength of the ambiguity effect.

For the MPT model, we present graphical posterior predictive checks for trial-type proportions separately for NP/Z sentences (Figure 1) and MV/RR sentences (Figure 2). The posterior predictive distribution is obtained by generating new data from the fitted model with each iteration of the sampling algorithm. Posterior predictive checks then serve to evaluate whether the model can reproduce the empirical patterns in the data (e.g., Gelman et al., 1996). Recall that there is no empirical trial-type data for the Fujita (2025) Maze experiment, as there was no end-of-trial task and rereading was impossible.

The figures show that the MPT model mostly replicates the observed qualitative and quantitative patterns, with the notable exception of the *accept+regression* and *reject+no regression* patterns for NP/Z sentences in K.-J. Huang et al. (2024). Figure 1 shows an empirically larger effect of ambiguity on the former pattern, while the model-generated data show a larger effect on the latter pattern. A similar but less pronounced mismatch is also visible for the second BSPR study of Paape and Vasishth (2022a). It is currently difficult to say where this mismatch is coming from, but it is clear that the MPT model overestimates the proportion of ambiguous sentences that are ultimately accepted/interpreted correctly after rereading, and underestimates the proportion of ambiguous sentences that are rejected/misinterpreted in the absence of rereading. We plan to investigate this pattern in future work.

Another surprising result is the overall higher proportion of predicted rejections/misinterpretations for the data of Fujita (2025), as well as a slight tendency for unambiguous sentences to be more rejected/misinterpreted *more* often than ambiguous sentences. As task information is missing from the Fujita (2025) data, these patterns must be driven exclusively by the reading times at the critical and spillover regions. We speculate that the result may be due to the $p\_postpone$ parameter: Maze is often argued to minimize or eliminate spillover (e.g., Boyce and Levy, 2023), and assuming the same spillover probability for SPR and Maze may lead the model to interpret fast reading times at the spillover region in Maze as evidence of inattention.

Table 1 shows the parameter estimates for the MPT model. In line with previous work, the estimates indicate a much higher incidence of garden-pathing in the ambiguous conditions compared to the unambiguous conditions, a relatively large cost of covert reanalysis, and more successful reanalysis in MV/RR compared to NP/Z sentences.

Table 1: Parameter estimates (back-transformed upper and lower bounds of 95% CrI) for the MPT model.

parameter	low	high	interpretation
$shift$	162 ms	170 ms	Non-decision time
att_cost	5 ms	7 ms	Cost of attention
gp_cost	11 ms	18 ms	Cost of garden-pathing
$reanalysis\_cost$	441 ms	494 ms	Covert reanalysis cost
$regression\_cost$	266 ms	332 ms	Regression cost (BSPR)
$p\_attentive$	0.90	0.92	Prob. of attentive reading
p_gp	0.18	0.20	Prob. of garden-pathing (unambiguous cond.)
p_overt	0.19	0.31	Prob. of overt reanalysis
$p\_postpone$	0.65	0.70	Prob. of postponing reanalysis to spillover
$p\_success\_o$	0.67	0.74	Success prob. of overt reanalysis
$p\_success\_c$	0.40	0.43	Success prob. of covert reanalysis
p_infer	0.23	0.37	Prob. of pragmatic inference
$p\_base\_regress$	0.01	0.02	Baseline regression prob. in BSPR
β_1	+0.42	+0.55	Garden-path prob. increase (ambiguous cond.)
β_2	+0.11	+0.20	Ambiguity effect MV/RR vs. NP/Z
β_3	+0.05	+0.18	Overt reanalysis success MV/RR vs. NP/Z
β_4	+0.22	+0.29	Covert reanalysis success MV/RR vs. NP/Z
β_5	+0.00	+0.01	Baseline regressions ET vs. BSPR

We carried out model comparisons via cross-validation with the loo R package (Vehtari et al., 2022). During cross-validation, a subset of the data is held out, and the model is fitted to the remaining data. After fitting, the expected log pointwise predictive density ($\widehat{elpd}$) of the held-out data is computed. This measure quantifies how well the fitted model predicts the previously unseen data points. The process is repeated multiple times for different subsets of the data. The model with the highest overall $\widehat{elpd}$ has the best predictive fit. One advantage over measures of in-sample fit (such as R^2) is that evaluating out-of-sample fit penalizes overfitting: a model that is too tightly optimized will generally not fit new data very well. A downside is that cross-validation is computationally expensive; the loo package overcomes this issue by computing an approximation via importance sampling (Vehtari et al., 2017). In general, moving away from optimizing in-sample fit towards optimizing out-of-sample fit is an important goal for cognitive science, as models with good out-of-sample fit should provide better generalizability (Yarkoni, 2022).

Table 2 shows the cross-validation results. The MPT model clearly outperforms the surprisal model in terms of predictive fit, and the MPT-plus-surprisal model slightly outperforms the base MPT model.

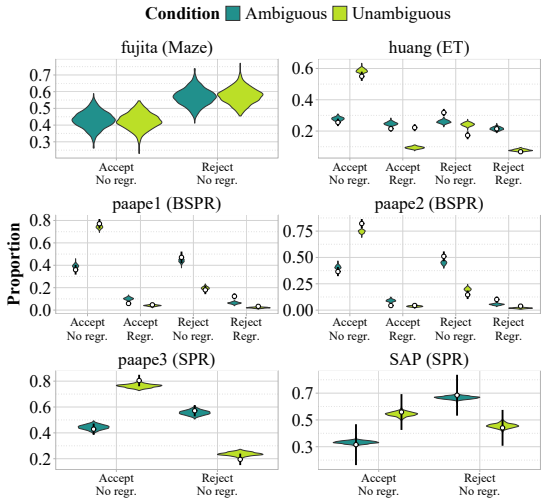

Figure 1: Observed (points/error bars) and predicted (violins) proportions of trial types for NP/Z sentences. Trials with incorrect question responses grouped as *reject* trials.

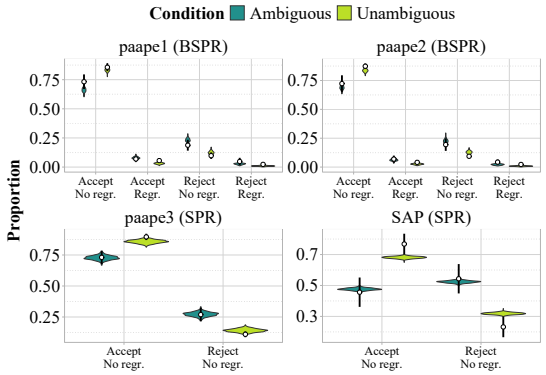

Figure 2: Observed (points/error bars) and predicted (violins) proportions of trial types for MV/RR sentences. Trials with incorrect question responses grouped as *reject* trials.

It is informative to ask which observations are driving the MPT model's predictive advantage, such as very short or long reading times (e.g., Nicenboim and Vasishth, 2018). Figure 3 plots the difference in $\widehat{elpd}$ between the base MPT and surprisal models against reading times at the critical sentence region by trial type. The MPT model tends to have an advantage for trials with regressions and for trials with longer reading times. Longer reading times being the driving force behind the MPT advantage is consistent with the assumption that the unique feature of garden-path sentences is a relatively large reanalysis cost (e.g., K.-J. Huang and Dillon, 2023).

Discussion

We have presented an application of the Paape and Vasishth (2022b) multinomial processing tree model of garden-pathing to data from four different reading paradigms: eye tracking, regular and bidirectional self-paced reading, and Maze. Reading times, rereading probabilities and end-of-trial question responses/grammaticality judgments are modeled in parallel, and the same underlying cognitive processes are assumed to drive participant's reading behavior across all four paradigms. The inclusion of a pragmatic inference process that explains away a subset of apparent misinterpretations allows for making a direct connection between question responses and grammaticality judgments. Jointly modeling different reading paradigms and end-of-sentence tasks not only increases the amount of data available for parameter estimation, but is also a step forward with regard to reconciling sometimes mismatching empirical findings with psycholinguistic theories that are usually formulated in a task-agnostic way (e.g., Witzel et al., 2012).

Our model outperformed GPT-2 surprisal in terms of predictive fit, in line with previous work that showed suboptimal fit of LLM surprisal to reading times in garden-path sentences (K.-J. Huang et al., 2024; Paape & Vasishth, 2025; Van Schijndel & Linzen, 2021). The result implies that humans may use specialized recovery mechanisms when processing these constructions that are unavailable to LLMs. That being said, adding surprisal to the MPT model did improve predictive fit, suggesting that next-word predictions do have a role to play in garden-path processing. Our work thus contributes to the ongoing discussion regarding the precise role of surprisal in sentence processing (e.g., Slaats and Martin, 2025; Staub, 2025), while also underscoring the need for more sophisticated mathematical models in reading research.

Table 2: Cross-validation results. Higher $\Delta\widehat{elpd}$ means better predictive fit compared to the null model.

Model	$\Delta\widehat{elpd}$ vs. null model	SE of $\Delta\widehat{elpd}$
MPT	9113.5	151.6
SURPRISAL	2191.4	66.8
MPT-PLUS-SURPRISAL	9452.5	154.0

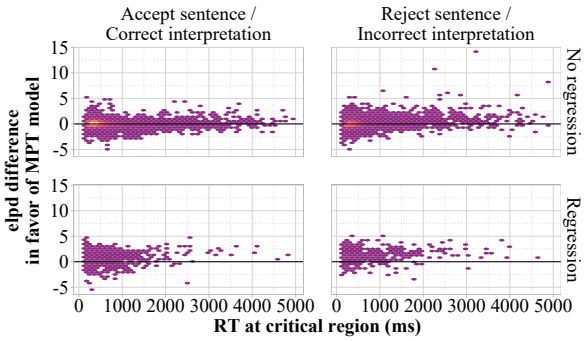

Figure 3: Pointwise $\Delta\widehat{elpd}$ between the base MPT and surprisal models by trial type and reading time. y-axis values above zero indicate data points that are better predicted by the MPT model. The predominance of positive values indicates an overall advantage for the MPT model over the surprisal model.

References

Boyce, V., & Levy, R. (2023). A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*, *2*(1).

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(3), 108–124.

Fodor, J. D., & Inoue, A. (2000). Garden path repair: Diagnosis and triage. *Language and Speech*, *43*(3), 261–271.

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

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies* [Doctoral dissertation]. University of Connecticut.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Fujita, H. (2025). The L-maze task and web-based data collection in second language sentence processing research. *Research Methods in Applied Linguistics*, *4*(3), 100251.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *cmdstanr: R Interface to 'CmdStan'* [https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org].

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 1–8.

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, *83*(4), 893–918.

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, *4*, 730570.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

Huang, K.-J., & Dillon, B. (2023). An infrequent, large cost underlies garden path effects: an RT distribution approach. *Talk presented at HSP 2023, University of Pittsburgh*.

Huang, Y., & Ferreira, F. (2021). What causes lingering misinterpretations of garden-path sentences: Incorrect syntactic representations or fallible memory processes? *Journal of Memory and Language*, *121*, 104288.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, *54*(4), 445–465.

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

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Lewis, R. L. (1998). Reanalysis and limited repair parsing: Leaping off the garden path. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 247–285). Kluwer.

Lopes Rego, A. T., Snell, J., & Meeter, M. (2024). Language models outperform cloze predictability in a cognitive model of reading. *PLOS Computational Biology*, *20*(9), e1012117.

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language*. MIT Press.

Meng, M., & Bader, M. (2000). Ungrammaticality detection and garden path strength: Evidence for serial parsing. *Language and Cognitive Processes*, *15*(6), 615–666.

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.

Paape, D., & Vasishth, S. (2022a). Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading. *Glossa Psycholinguistics*, *1*(1).

Paape, D., & Vasishth, S. (2022b). Estimating the true cost of garden pathing: A computational model of latent cognitive processes. *Cognitive Science*, *46*(8), e13186.

Paape, D., & Vasishth, S. (2022c). Is reanalysis selective when regressions are consciously controlled? *Glossa Psycholinguistics*, *1*(1).

Paape, D., & Vasishth, S. (2025). Context ameliorates but does not eliminate garden-pathing: Novel insights from latent-process modeling [OSF preprint]. https://osf.io/preprints/osf/amcnq_v2

Pritchett, B. L. (1992). *Grammatical competence and parsing performance*. University of Chicago Press.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121.

Slaats, S., & Martin, A. E. (2025). What's surprising about surprisal. *Computational Brain & Behavior*, *8*(2), 233–248.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Stan Development Team. (2024). Stan reference manual, version 2.36.0.

Staub, A. (2025). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*(1), 17–34.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. The R Foundation. https://mc-stan.org/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Warner, J., & Glass, A. L. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, *26*(6), 714–738.

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.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1.