# A-maze of Natural Stories: Texts are comprehensible during the Maze task

Veronica Boyce[1] & Roger Levy[2]

[1] Stanford University
[2] Massachusetts Institute of Technology

## Author Note

TODO

The authors made the following contributions. Veronica Boyce: Conceptualization, Formal Analysis, Investigation, Software, Writing - Original Draft Preparation, Writing - Review & Editing; Roger Levy: Conceptualization, Formal Analysis, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Veronica Boyce, 450 Jane Stanford Way, Building 420, Stanford University, Stanford, CA 94305. E-mail: vboyce@stanford.edu

## Abstract

We find support for the localization of reading time effects during the Maze task, as well as extending the range of materials Maze is suitable for. How long it takes to read a word in a sentence is reflective of how hard it is to identify and integrate the word in the surrounding context. Techniques that slow down the reading process and localize the processing time for each word are useful to understanding the time course of language processing. A-maze is a new method for measuring incremental sentence processing that can localize slowdowns related to syntactic ambiguities. We adapt A-maze for use on longer passages and test it on the Natural Stories corpus. We find that people can comprehend what they read during the Maze task. Moreover, the task yields useable reaction time data with word predictability effects that are linear in the surprisal of the current word, with little spillover effect from the surprisal of the previous word. This expands the types of effects that can be studied with A-maze, showing it to be a a versatile alternative to eye-tracking and self-paced reading.

*Keywords:* TODO
Word count: X

# A-maze of Natural Stories: Texts are comprehensible during the Maze task

## Intro

It's remarkable how flexible we are when reading; while we do occasionally stumble when we read something unexpected, we often are able to read slightly unexpected things without a problem. However, these expectations shape how fast we read, even if we don't notice a stumble, unexpected words take longer to process as they force us to rebuild our burgeoning mental model of the sentence. Fortunately for fluent readers and unfortunately for studying language, this process of adjustment is very quick, which makes measures of reading time messy.

Measures of online reading are one way to understand language and how the mind processes language. Many theories of language structure and language processing ground out in predictions about the difficulty of processing words. For instance, the subject v object relative debate includes theories that make fine-grained predictions about which word is how slow – this needs localized methods to adjudicate it. Other theories such as noisy channel processing also need support from localized word-by-word results. TODO WHY *DO* we care about localization?

Incremental processing methods such as self-paced reading or eye-tracking measure how long someone spends looking at one word before moving on and use that as a proxy for how difficult word was in context. Across multiple methods for measuring per-word processing and reading time, how unexpected a word is correlates with how long it takes to read it and move on (CITE). However, the two major methods of measuring incremental processing both suffer from imprecise localization. In eye-tracking, people read naturally which involves skipping words, jumping ahead and looking back, the dynamics of which make it hard to isolate effects. Even when reading order is controlled in self-paced reading, readers are moving quickly through the text and it may take multiple words for slowdowns to catch up with them.

One effect of this lack of localization is that reading time for a word is dependent not only on how unexpected a word is, but also how unexpected the previous word is. This is an indication of spillover from the previous word.

It's well established for eye-tracking and SPR that RTs are roughly linear in terms of a word's surprisal (negative log probability). Due to spillover effects, on SPR and eye-tracking, there is also a positive linear relationship between the surprisal of a previous word and the RT on the current word – this is an indication of lack of localization. In addition to suprisal predicting RT, word length and word's overall frequency are also often found to be predictive. TODO many CITES

An alternative method that seems to have superior localization is the Maze task, which adopts an unnatural way of reading to force incremental processing (Forster, Guerrera, & Elliot, 2009). In the Maze task, participants see two words at a time, a correct word that continues the sentence, and a distractor which does not. Participants must choose the correct word, and their reaction time (RT) is the dependent measure. If participants make a mistake, the sentence discontinues. Theoretically, participants must fully integrate each word into the sentence in order to confidently select it. This idea is supported by studies finding strongly localized effects (**???**).

The downside of Maze is that materials are effort-intensive to construct because of the need to select infelicitious words as distractors for each spot of each sentence; this may explain why the Maze task was not widely adopted. Boyce, Futrell, and Levy (2020) demonstrate a way to automatically generate Maze distractors by using NLP language models to find words that are high-surprisal in the context of the target sentence. The quality of these A-Maze distractors is not up to that of hand-generated distractors, but Boyce et al. (2020) found that materials with A-maze distractors had similar results to the hand-generated distractors from (**???**). A-maze out performed a SPR control in detecting and localizing expected slowdown effects. Sloggett, Handel, and Rysling (n.d.) also found that A-maze and G-maze distractors yielded similar results on a disambiguation paradigm.

A-maze is a potentially powerful addition to the psycholinguists toolkit. However, like the other Maze tasks it has been limited in its application to single-sentence items probing minimal comparisons in constructed sentences. This limits its useful, as some important questions such as comparing human data to language models or studying discourse effects require processing times over multi-sentence passages. Therefore, it's important to expand the Maze task to these types of materials and verify that it finds comparable patterns to other methods.

While the issue of needing to generate distractors for long passages is solved with A-maze, another problem with Maze remains. In particular, because Maze tasks discontinue after participants make mistakes, the farther into an item a word is, the fewer participants see it. This makes it hard to run long materials using Maze, and prevents Maze from being used on long text passages where SPR or eye-tracking could be used.

We address this problem by creating a Maze task interface where participants correct their mistakes and continue with the sentence rather than terminating on mistakes. This allows for multi-sentence items to be run using Maze, which we verify by running Maze on the Natural Stories corpus. With this data, we are able to confirm that participants can comprehend what they read using Maze and investigate the effects of surprisal on RT in Maze.

## Error-correction maze

One advantage of the Maze task is that it forces incremental processing and automatically excludes inattentive participants by terminating a sentence when a participant makes an error. Thus, only participants who have processed the sentence and are paying attention contribute data at critical regions later in the sentence. However, this means we don't have data after a participant makes a mistake in an item. In traditional G-maze tasks, with hand-crafted distractors and attentive participants, this is a small issue. However, this data loss is much worse with A-maze materials and crowd-sourced participants (Boyce et al.,

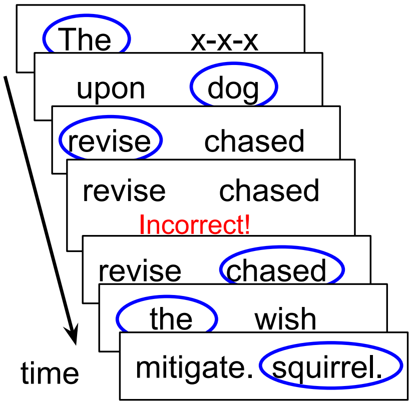

*Figure 1*. Schematic of error-correction Maze. A participant reads a sentence word by word, choosing the correct word at each time point (selections marked in blue ovals). When they make a mistake, an error message is displayed, so they try again and continue with the sentence.

2020). The high errors are likely from some combination of participants guessing randomly and from auto-generated distractors that in fact fit the sentence; as Boyce et al. (2020) noted, some distractors, especially early in the sentence, were problematic and caused considerable data loss.

This situation could be improved by auto-generating better distractors or hand-replacing problematic ones, but that does not solve the fundamental problem. Well-chosen distractors and attentive participants will reduce the error rate, but the error rate will still compound over long materials. For instance, with a 1% error rate, 86% would complete each 15-word sentence, but only 61% of participants would complete a 50 word vignette, and 13% a 200 word passages. In order to run longer materials, we need something to do when participants make a mistake, other than terminate the entire item.

To resolve this, we introduce an *error-correction* variant of Maze shown in Figure 1. When a participant makes an error, we present them with an error message and wait until they select the correct option, before continuing the sentence as normal. We make this "error-correction" Maze available as an option in a modification of the Ibex Maze implementation introduced in Boyce et al. (2020) (https://github.com/vboyce/Ibex-with-Maze). The code records both the RT to the first click and also the total RT until the correct answer is selected as separate values.

This variant of Maze expands the types of materials that can be used with maze to include arbitrarily long passages and cushions the impact of occasional problematic distractors.

## Natural Stories corpus

We test this error-correction Maze on the Natural Stories corpus. The Natural Stories corpus (Futrell et al., 2020) consists of 10 passages each roughly 1000 words long which are designed to read fluently to native speakers. At the same time, the passages contain copious punctuation, quoted speech, many proper nouns, and low frequency grammatical constructions. Taken together, these properties make this a severe test of our process, as these features make it harder for the language model to choose good distractors and require focus from participants. If participants can succeed at the Maze task on this set of materials, we think they are likely to succeed on basically any naturalistic text.

The corpus is accompanied by binary-choice comprehension questions, 6 per story, which we use to assess comprehension. Using this corpus on the A-maze task allows us to address, first whether participants will read and understand long passages using A-maze with error correction, and second, whether the resulting RTs profiles will show expected patterns such as a linear relationship with surprisal.

## Methods

We constructed A-maze distractors for the Natural Stories corpus (Futrell et al., 2020) and recruited 100 crowd-sourced participants to each read a story in the Maze paradigm.

### Materials

We took the texts of the Natural Stories corpus (Futrell et al., 2020), split them into sentences, and ran the sentences through the A-maze generation process. We follow the A-maze generation process outlined in Boyce et al. (2020), although we use an updated version of the codebase with fixes some issues identified in that paper (code at https://github.com/vboyce/Maze). Additionally, the capability to appropriately handle a wider variety of punctuation (needed for this corpus) was added. We took the auto-generated distractors as they were, without checking them for quality. We used the original comprehension questions provided in the Natural Stories corpus. To familiarize participants with the task, we wrote a short practice passage and corresponding comprehension questions. All materials are available at TODO NEW REPO.

### Participants

We recruited 100 participants from Amazon Mechanical Turk, and paid each participant 3.50 dollars, for roughly 20 minutes of work. We excluded data from those who did not report English as their native language, leaving 95 participants.

### Procedure

Participants first gave their informed consent and saw task instructions. Then they read a short practice story in the Maze paradigm and answered 2 binary-choice practice comprehension questions, before reading the main story in the A-maze task. After the story, they answered the 6 main comprehension questions, commented on their experience, answered optional demographic questions, saw an debriefing and were given a code to enter for payment. The experiment was implemented in Ibex (https://github.com/addrummond/ibex) and the experimental code is available at REPO.

## Data analysis

We conducted data processing and analyses using R (Version 4.0.3; R Core Team, 2020) and the R-packages *brms* (Version 2.14.4; Bürkner, 2017, 2018), *cowplot* (Version 1.1.1; Wilke, 2020), *gridExtra* (Version 2.3; Auguie, 2017), *here* (Version 1.0.1; Müller, 2020), *lme4* (Version 1.1.26; Bates, Mächler, Bolker, & Walker, 2015), *mgcv* (Version 1.8.33; Wood, 2011, 2003, 2004; Wood, N., Pya, & S"afken, 2016), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *patchwork* (Version 1.1.1; Pedersen, 2020), *tidybayes* (Version 2.3.1; Kay, 2020), and *tidyverse* (Version 1.3.0; Wickham et al., 2019).

To model the relationship between RT and word surprisal, we created a set of predictor variables of frequency, word length, and surprisals from three language models. For length, we used the length in characters excluding end punctuation. For unigram frequency, we tokenized the training data from Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018) and tallied up instances. We then used the log2 frequency of the expected occurances in 1 billion words as the model predictor, so higher values indicate higher log frequencies. We got per-word surprisals for each of 3 different language models: a Kneser-Ney smoothed 5-gram, GRNN (**???**), and Transformer-XL (**???**). For all of these predictors, we consider both the predictor at the current word as well as lagged predictors from the previous word.

We only included words that were a single token in each of the model vocabularies and for which we had frequency information. In practice, this excluded words with punctuation as well as uncommon or proper nouns. We also excluded the first word of every sentence (which had a dummy distractor). We excluded outlier RTs that were <100 or >5000 ms (<100 is likely a recording error, >5000 is likely the participant getting distracted). We exclude words where mistakes occurred or which occurred after a mistake in the same sentence.

For generalized additive models, we centered but did not rescale the length and frequency predictors, but left surprisal uncentered for interpretability. We used smooths for the surprisal terms and tensor effects for the frequency by length effects and interactions.

For linear models, we centered all predictors. We used full mixed effects, including by-subject slopes and a per-word random intercept. We used weak priors (normal(1000,1000) for intercept, normal(0,500) for beta and sd, and lkj(1) for correlations). Models were run in BRM.

For model comparison, we fit models with only frequency and length as predictors, as well as models that also had one or more sources of surprisal. We centered all effects.

**TODO SPR!!!**

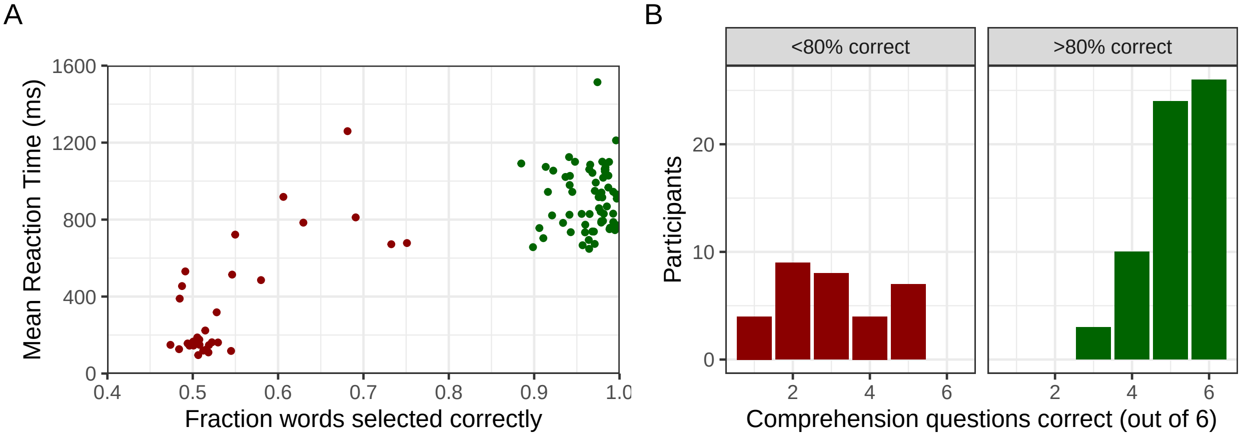

*Figure 2*. A. Correlation between a participant's accuracy on the Maze task (fraction of words selected correctly) and their average reaction time (in ms). Many participants (marked in green) chose the correct word >80% of the time; others (in red) appear to be randomly guessing and were excluded from further analysis. B. Performance on the comprehension questions. Participants who had >80% task accuracy tended to do well; those who were at chance on the task were also at chance on the questions.

# Results

## Reading stories in the Maze task

Many participants completed the Maze task with a high degree of accuracy and also answers the comprehension questions correctly.

Participant accuracy reflects both how well participants can navigate the task and what quality the auto-generated distractors are. We calculated the per-word error rate for each participant and graphed it against their average reaction time. (To avoid biasing the average if a participant took a pause before returning to the task, RTs greater than 5 seconds were excluded.) As seen in Figure 2A, one cluster of participants (marked in green) make relatively few errors, with some reaching 99% accuracy. This confirms that the distractors were generally appropriate and shows that some participants maintained focus on the task for the whole story. These careful participants took around 1 second for each word selection, which is much slower than other paradigms CITE. Another cluster of participants (in red) sped through the task, seemingly clicking randomly. This bimodal distribution is likely due to the mix of workers on Mechanical Turk, as we did not use qualification cutoffs.

Another check is whether participants comprehended the story. We counted how many of the binary-choice comprehension questions each participant got right (out of 6). As seen in Figure 2B, most participants were were accurate on the task also did well on comprehension questions, while participants who were at chance on the Maze task were also at chance on the comprehension questions. Participants usually answered quickly (within 10 seconds), so we do not believe they were looking up the answers on the internet. We can't rule out that some participants may have been able to guess the answers without understanding the story. Nonetheless, this provides preliminary evidence that people can

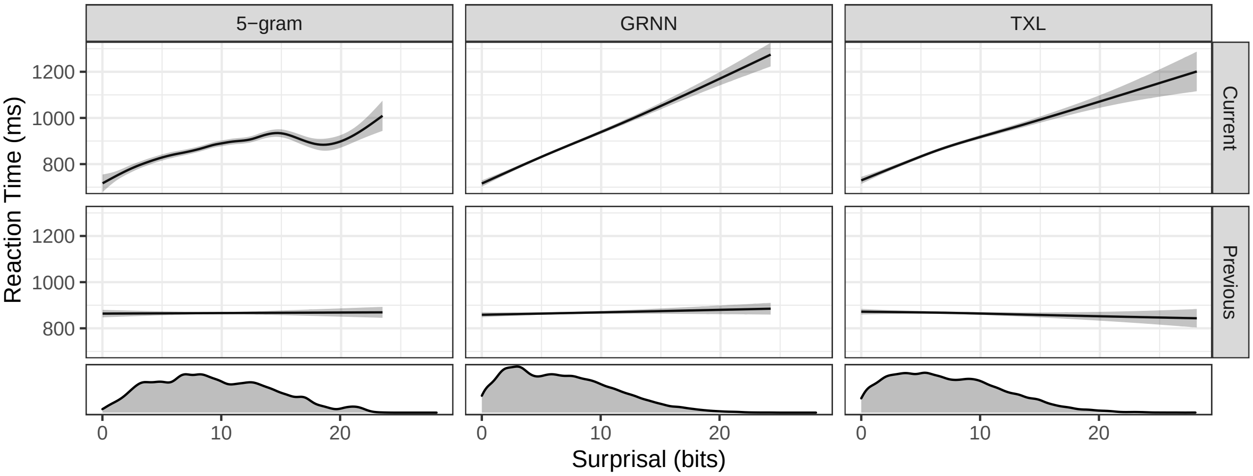

*Figure 3*. GAM predictions of reaction time (RT) as a function of either current word surprisal (top) or previous word surprisal (bottom). Density of data is shown along the x-axis. For each of the 3 language models used, there is a linear relationship between current word surprisal and RT (at least when there is enough data). There is no relationship between previous word surprisal and RT.

understand and remember details of stories they read during the Maze task.

We use task performance as our exclusion metric and only analyze data from participants with at least 80% accuracy (in the gap between high-performers and low-performers).

## RT and surprisal

We fitted generalized additive models to test whether the RTs from the Maze experiment showed a linear relationship with surprisal and whether the effect was limited the current word or had spilled over from the prior word. For these models, we only included data that occurred before any mistakes in the sentence; due to limits of model vocabulary, words with punctuation and some uncommon or proper nouns were excluded.

The smooths for the current and previous words surprisals are shown in Figure 3. Note that for each of the models, high-surprisal words are rare, with much of the data for words between 0 and 15 bits of surprisal. All of the models show a roughly linear relationship between current word surprisal and RT, especially in the region with more data. All of the models show a flat relationship between previous word surprisal and RT. This is a sign of localization as the previous word's surprisal is not affecting RT, only the word's own suprisal is. The linear relationship matches that found with other methodologies.

**TODO comparison with SPR !!!!** Given that the GAM models show a roughly linear relationship, we fit mixed linear models to quantify the influence of surprisal, frequency and word length. We built linear models with surprisal, frequency, length and surprisal x length and frequency x length effects from the current and previous words as predictors.

As we can see in Table 1, we find large effects of surprisal and length, but minimal

Table 1

*Predictions from fitted Bayesian regression models. All terms were centered, but not rescaled. Units are in ms. Surprisal is per bit, length per character, and frequency per* $log_2$ *occurance per billion words.*

| Term | 5-gram | GRNN | TXL |
|---|---|---|---|
| Intercept | 865.3 [829.9, 902.9] | 871.1 [837.9, 905.3] | 870.8 [832.5, 907.8] |
| Surprisal | 11.7 [9.3, 14.1] | 23.7 [21, 26.5] | 18.5 [16.1, 21.1] |
| Length | 20.5 [15.4, 25.6] | 18.5 [13.3, 23.7] | 21.4 [16.2, 26.6] |
| Frequency | -2.9 [-6.3, 0.5] | 2.9 [-0.2, 6] | 0.4 [-2.7, 3.5] |
| Surp x Length | -2 [-3, -1] | -1.8 [-2.7, -0.9] | -1.4 [-2.2, -0.6] |
| Freq x Length | -1 [-2.5, 0.4] | -0.1 [-1.2, 1] | 0.2 [-0.9, 1.2] |
| Past Surprisal | 1.6 [-0.5, 3.6] | 2.7 [0.8, 4.5] | 0.8 [-0.9, 2.5] |
| Past Length | -4.8 [-9, -0.1] | -6.6 [-10.9, -2.1] | -5.2 [-9.3, -0.7] |
| Past Freq | 2.6 [-0.1, 5.4] | 1.9 [-0.2, 4.2] | 1.2 [-1.1, 3.6] |
| Past Surp x Length | -0.2 [-1.2, 0.8] | -0.9 [-1.7, -0.2] | -0.6 [-1.3, 0.2] |
| Past Freq x Length | -1 [-2.3, 0.3] | -1.8 [-2.9, -0.8] | -1.5 [-2.6, -0.5] |

effects of frequency. These effects are larger than what is usually reported in other methods CITE, but this could be due to the overall slowness of the method. The lack of frequency effects is somewhat surprising, but consistent with CITE shain. Notably the coefficients for the lagged terms are small relative to the effects of surprisal and length of the current word.

As a last analysis, we checked which of our surprisal models had the best fit using a nested model comparison. We found that all surprisal sources provide predictive value over none, but that the information provided by the Ngram model does not provide additional value to a model that already has GRNN in it. TXL and GRNN appear to contain some complementary predictive value.

## Discussion

Between auto-generating Maze distractors and the error-correction paradigm, Maze is a good complement to existing incremental processing methods. It localizes effects better that SPR does while showing similar patterns in terms of surprisal. In additon, despite the oddness of the task participants can succeed at it while understanding what they are reading. [summarize results] - error-correction paradigm - test on natural stories - it works

While there are some differences, especially in scale, to that found with other incremental processing paradigms, we think this is a reason to explore more. Comparisons between methods could be very useful for identifying how task demands influence processing.

All the code is available, and we encourage researchers to consider trying out Maze if they think it's appropriate to their experiments. This opens up another area of research to being amenable to the new A-maze paradigm.

The Maze task might be espeically good for comparing with NNs because of the forced incrementality. While eventually we do want to figure out models of more natural reading, this might be a useful complement b/c we don't have to deal with nasty spillover.

Secondly, this error-correction compensates for some of the shortcomings of A-Maze identified in Boyce et al. (2020); distractors might still cause participants to make unavoidable mistakes, but at least the still see the rest of the sentence and we get their data. Whether this also solves the data loss problem from the researcher perspective within a sentence depends on whether post-mistake data are high-quality and trustworthy; this is a hard-to-assess question of potential interest.

Even if researchers exclude all post-mistake data from analysis, this process can still address the question of whether errors are due to inattentive participants or bad distractors (as discussed in Boyce et al. (2020), section XX). When we have a participants selection for every word, we can easy calculate a per-word error rate for each participant, without having to address the censoring present in error-terminating Maze (where we can't know if a participant would have made more mistakes after the first; with error correction, we know how may more they did make). If participants are guessing, we'll see high per-word error rates, if the early distractors are bad, we'll see low per-word error rates (with errors concentrated at the start of the sentences). This per-word error rate metric measures participants task accuracy, and allows us to exclude data from inattentive participants. This provides a convenient and clear-cut way of controlling data quality after the fact.

It also starts to reduce the perverse incentives. With error-terminates Maze, the fastest way to get through the task is to select randomly, and it's quite quick because mistakes skip you to the next sentence. With error-correction Maze, randomly jamming buttons takes more effort, but is still an effective strategy. [Footnote: In discussing this work, we recieved the suggestion that one way to disincentivize random clicking is to add a pause when a participant makes a mistake, forcing them to wait some short period of time (ex 500ms or 1 sec) before being able to correct their mistake. This seems like a promising improvement that could be worth implementing and testing.]

## References

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze Made Easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082. https://doi.org/10.1016/j.jml.2019.104082

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

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

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2020). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Lang Resources & Evaluation*. https://doi.org/10.1007/s10579-020-09503-7

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

Kay, M. (2020). *tidybayes: Tidy data and geoms for Bayesian models*. https://doi.org/10.5281/zenodo.1308151

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

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

Sloggett, S., Handel, N. V., & Rysling, A. (n.d.). A-maze by any other name, 1.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=cowplot

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, *65*(1), 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wood, S. N., N., Pya, & S"afken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, *111*, 1548–1575.