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 these longer text passages 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: TODO

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 (Bartek, Lewis, Vasishth, & Smith, 2011). 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 (Grodner & Gibson, 2005; Staub, 2010; Traxler, Morris, & Seely, 2002). Other theories such as noisy channel processing also need support from localized word-by-word results (Levy, 2011).

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. How unexpected a word is correlates with how long it takes to read it and move on (Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner, Ashby, Pollatsek, & Reichle, 2004). 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 (Frazier & Rayner, 1982; Levy, Bicknell, Slattery, & Rayner, 2009; Rayner, 1998; Rayner et al., 2004). Even when reading order is controlled in self-paced reading, readers may maintain ambiguities about what a word was or means until it is later resolved by context, so it may take multiple words for slowdowns to catch up with them (Koornneef & van Berkum, 2006; MacDonald, 1993).

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) (Goodkind & Bicknell, 2018; Luke & Christianson, 2016; Smith & Levy, 2013; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). 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 (Kliegl et al., 2004).

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; Freedman & Forster, 1985). 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. Two versions of the maze task exist: lexical or L-maze where the distractors are nonce words and grammatical or G-maze where the distractors are real words of the language that don't fit in the context of the sentence so far. 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 for G-maze which is more sensitive than L-maze (Witzel, Witzel, & Forster, 2012).

The downside of G-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 G-maze distractors, but Boyce et al. (2020) found that materials with A-maze distractors had similar results to the hand-generated distractors from Witzel et al. (2012). A-maze outperformed L-maze and an SPR control in detecting and localizing expected slowdown effects. Sloggett, Handel, and Rysling (2020) 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 resolve this issue by introducing a new paradigm for running Maze tasks where participants correct their errors and continue reading rather than discontinuing after a mistake.

Using this tweak, we test A-maze on the passages from 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 a wide variety of naturalistic texts. The corpus is accompanied by binary-choice comprehension questions, 6 per story, which we use to assess comprehension. Futrell et al. (2020) collected self-paced reading time data on this corpus, which we compare the A-maze results to. Participants were able to read and understand these long passages using A-maze with the error correction paradigm, and their RT profiles showed a linear relationship to the surprisal of the words.

Error-correction maze

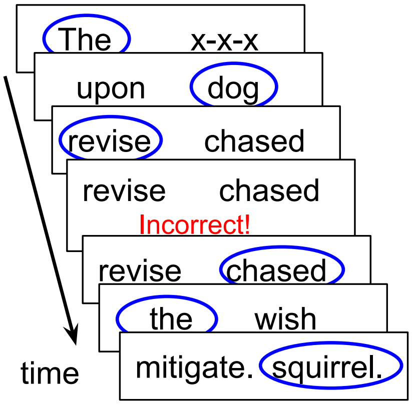

Figure 1. Schematic of error-correction Maze. A participant reads a sentence word by word, choosing the correct word at each time point (selections marked in blue ovals). When they make a mistake, an error message is displayed, so they try again and continue with the sentence.

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., 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.

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. See the Appendix for an except of one of the stories and its 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 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.

Models

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), *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), *tidybayes* (Version 2.3.1; Kay, 2020), and *tidyverse* (Version 1.3.1; 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 4 different language models: a Kneser-Ney smoothed 5-gram, GRNN (Gulordava et al., 2018), Transformer-XL (Dai et al., 2019), and GPT-2 (Radford et al., n.d.), using lm-zoo (Gauthier, Hu, Wilcox, Qian, & Levy, 2020).

We exclude the first word of every sentence because it has a different distractor, which leaves 9782 words. Then we exclude words for which we don't have surprisal or frequency information, leaving 8489 words. We additionally exclude words that were treated as being composed of multiple tokens. This is primarily words that have punctuation, where the punctuation is treated as a separate token. While surprisals should be additive, the summing the surprisals over these parts gives unreasonable responses sometimes. For instance, in one story the word king!' is given a surprisal of 64 by GRNN (context: The other birds gave out one by one and when the eagle saw this he thought, "What is the use of flying any higher? This victory is in the bag and I am king!"). To avoid these outliers we exclude all words that a model treated as multi-token, leaving 7512 words. (While GPT-2 using byte-pair encoding that can split up words into multiple parts, excluding words it split up only excludes 30 words that were not already excluded by other models.)

For all of these predictors, we consider both the predictor at the current word as well as lagged predictors from the previous word. 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. We only analysed words where we had values for all predictors, which means that if the previous word was unknown to a model, this word will be excluded because we're missing values for a lagged predictor.

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. To minimize the effect of repeated measures on confidence about inferred curve shapes, we collapse the reading times across subjects by modelling the mean RT for each word.

For linear models, we centered all predictors. We used full mixed effects, including by-subject slopes and a per-word-token random intercept (Barr, Levy, Scheepers, & Tily, 2013). 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 (Bürkner, 2018).

For model comparison, we took by-item averaged data to aid in fast model fitting. We included frequency, length, and their interaction to all models. Then we fit models with either 1 or 2 sources of surprisal using lm and assessed the effect of adding the second surprisal source with an anova. We used predictors for the current and past word and centered all effects.

Self-paced reading comparison

In addition to the texts, Futrell et al. (2020) released reading time data from a SPR study they ran in 2011. They recruited 181 participants from Amazon Mechanical Turk, most of whom read 5 of the stories. After reading each story, each participant answered 6 binary-choice comprehension questions. As a comparison to our A-maze models, we run similar models on the SPR corpus on Natural Stories (Futrell et al., 2020). For comparability, we analyse only the first story each participant read, and, in line with Futrell et al. (2020), exclude participants who got less than 5/6 of the comprehension questions correct. To account for spill over effects known to exist in SPR, we analyse predictors at the current word as well as the past 3 words for all models. For linear models, we centered all predictors. We were unable to fit the full mixed effects model. The best model we could fit had by-subject random intercept, uncorrelated by-subject random slopes for surprisal, length and frequency, and a per-word-token random intercept, fit with lme4, as this structure did not fit reliably in brms.

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. 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.

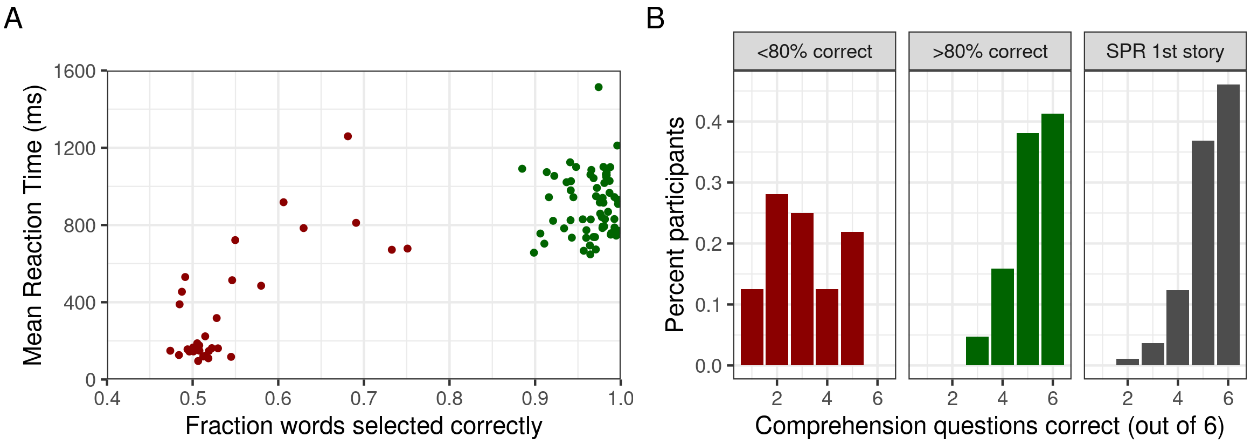

Figure 2. A. Correlation between a participant's accuracy on the Maze task (fraction of words selected correctly) and their average reaction time (in ms). Many participants (marked in green) chose the correct word >80% of the time; others (in red) appear to be randomly guessing and were excluded from further analysis. B. Performance on the comprehension questions. Participants with low accuracy also performed poorly on comprehension questions; Participants with >80% task accuracy tended to do well; their performance was roughly comparable to the performance of SPR participants from Futrell et al. (2020) on their first stories.

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 understand and remember details of stories they read during the Maze task. The comprehension question performance of accurate Maze participants is broadly similar to the performance of SPR participants from Futrell et al. (2020) on the first story read. Overall, 60% of Maze participants got 5 or 6 questions right (22% of low-accuracy participants and 79% of high-accuracy participants) compared to 91% of all SPR reads and 83% of 1st SPR reads. Note that these differences cannot be directly attributed to methods, as the participant populations differed. While both studies were conducted on Mturk, the quality of Mturk data has decreased from 2011 when the SPR was collected to 2020 when the A-maze was collected.

We use task performance as our exclusion metric because it is more fine-grained and only analyze data from participants with at least 80% accuracy (in the gap between high-performers and low-performers). For the SPR comparison, we follow Futrell et al. (2020)'s criteria and exclude participants who got less than 5 of the questions correct.

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.

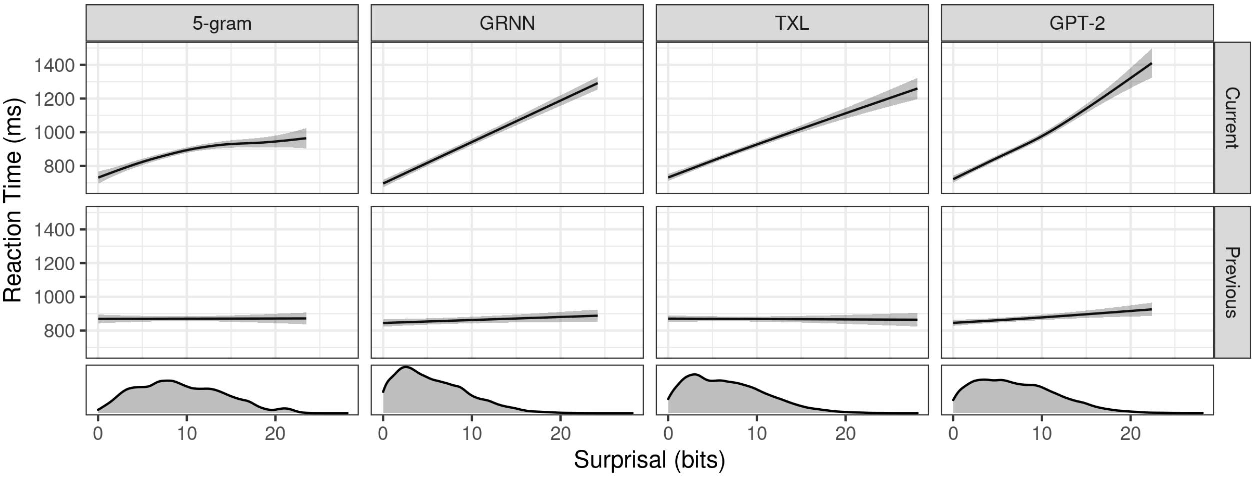

Figure 3. GAM predictions of reaction time (RT) as a function of either current word surprisal (top) or previous word surprisal (bottom). Density of data is shown along the x-axis. For each of the 4 language models used, there is a linear relationship between current word surprisal and RT (at least when there is enough data). There is no relationship between previous word surprisal and RT.

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.

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. We centered, but did not rescale, the predictors.

As we can see in Table 1, we find large effects of surprisal and length, but minimal 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 Shain (2019). Notably the coefficients for the lagged terms are small relative to the effects of surprisal and length of the current word.

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	TXL	GRNN	GPT-2
Intercept	876 [840.4, 910.9]	880 [842.8, 914.9]	876.8 [840.1, 911.5]	878.5 [845.6, 911.6]
Surprisal	11.1 [8.7, 13.6]	17.8 [15.3, 20.2]	22.3 [19.7, 25]	24.2 [21.5, 27]
Length	21.4 [16.6, 26.3]	20.5 [15.6, 25.4]	17.9 [13.2, 22.7]	16.2 [11.3, 21.2]
Frequency	-3.2 [-6.7, 0.5]	-0.1 [-3.2, 2.9]	1.8 [-1.1, 4.7]	-1.4 [-4.2, 1.2]
Surp x Length	-2 [-3, -0.9]	-1.4 [-2.1, -0.6]	-2.1 [-3, -1.2]	-1.8 [-2.7, -1]
Freq x Length	-1 [-2.5, 0.6]	0.1 [-1, 1.1]	-0.4 [-1.5, 0.7]	0.1 [-0.9, 1.1]
Past Surprisal	1.5 [-0.6, 3.5]	0.9 [-0.7, 2.5]	2.7 [1, 4.4]	3.5 [1.8, 5.3]
Past Length	-3.5 [-7.8, 0.7]	-3.7 [-7.7, 0.3]	-4.8 [-9, -0.8]	-5.1 [-9.2, -1.1]
Past Freq	2.5 [-0.3, 5.4]	1 [-1.3, 3.3]	1.8 [-0.4, 4]	0.7 [-1.4, 2.8]
Past Surp x Length	-0.2 [-1.1, 0.8]	-0.5 [-1.2, 0.2]	-0.9 [-1.7, -0.2]	-1.1 [-1.8, -0.4]
Past Freq x Length	-1 [-2.4, 0.4]	-1.5 [-2.5, -0.4]	-1.8 [-2.8, -0.8]	-1.7 [-2.7, -0.8]

As a last analysis, we checked which of our surprisal models had the best fit using a nested model comparison shown in Table 2. We assess the benefits of adding each model's predictions as a second surprisal source to see which models pick up on information not contained in another model. In terms of log likelihoods, GPT-2 is better than GRNN is better than TXL is better than 5-gram. Model comparisons tell a similar story; GPT-2 provides a lot of additional predictive value over each other model, GRNN provides a lot over 5-gram and TXL and a little complemetary information over GPT-2. TXL provides a lot over 5-gram, and 5-gram provides little over any model.

Table 2

Results of model comparisons on Maze data. Each row shows the additional predictive value gained from adding that model to another model. F values and p values from anova tests are reported. We also report log likelihoods of models with only one surprisal source.

Model	over 5-gram	over GRNN	over TXL	over GPT-2	Log Lik	r_squared
5-gram		2 (p=0.153)	3 (p=0.035)	0 (p=0.611)	-43817	0.16
GRNN	287 (p<0.001)		113 (p<0.001)	13 (p<0.001)	-43544	0.23
TXL	174 (p<0.001)	5 (p=0.006)		2 (p=0.137)	-43650	0.2
GPT-2	394 (p<0.001)	113 (p<0.001)	213 (p<0.001)		-43445	0.25

Comparison with SPR

As a comparison, we ran the same types of models on the Self-Paced Reading data collected by Futrell et al. (2020). As shown in Figure 4, there is a roughly linear but fairly flat relationship between RT and surprisal. Note that the y-axis is fairly narrow, and so the predicted effect of changes in surprisal is fairly small. This is confirmed by linear models (see Table 3). Surprisal and length effects are evident for the current word, and most models

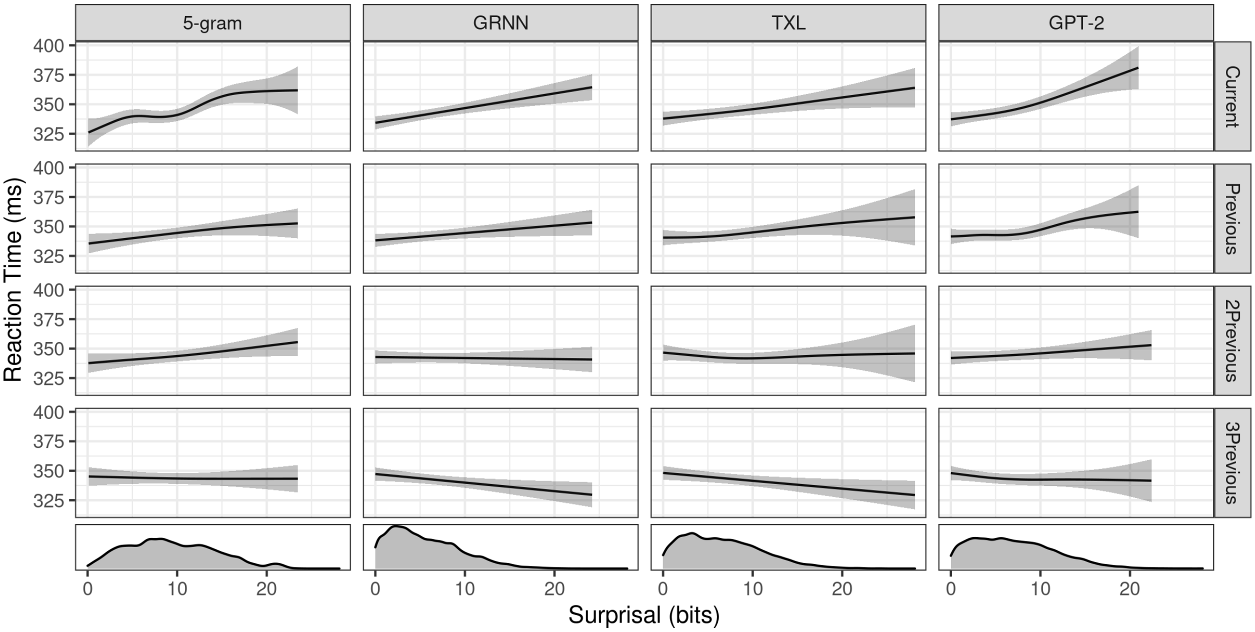

Figure 4. GAM predictions of reaction time (RT) for SPR data from Futrell et al. (2020) as a function of current word surprisal (top) or the surprisal of an earlier word, up to 3 words back. Density of data is shown along the x-axis.

shows surprisal effects from the past word, but we do not see frequency effects. The effects are much smaller than for A-maze, even though this model accounts for spillover effects from more previous words.

Similarly to the model comparison for Maze, we also conducted a model comparison on the SPR data shown in Table 4. In terms of log likelihoods, we find that GPT-2 is better than 5-gram is better than GRNN is better than TXL, although differences are small. Model comparisons show that GPT-2 and 5-gram models contain some value over each other model, which is less clear for TXL and GRNN. The relatively good fit of 5-gram models to SPR data compared with neural models matches results from Hu, Gauthier, Qian, Wilcox, and Levy (2020) and Wilcox et al. (2020). This contrasts with the Maze results, where the 5-gram model has the worst fit and does not provide additional predictive value to the other models.

As an overall measure of fit to data, we calculate multiple R-squared for the single surprisal source models for both A-maze and SPR. The models predict A-maze better with R-squared values ranging from 0.16 for the 5-gram model to 0.25 for GPT-2. Whereas for SPR, the R-squared values range from from 0.01 to 0.02. This suggests that the effect size differences are not due merely to the larger overall reading time for A-maze, but that instead A-maze is more sensitive to surprisal and length effects.

Discussion

We introduced a tweak to the paradigm for displaying Maze tasks that makes it useable for multi-sentence materials. In this error-correction Maze paradigm, participants

Table 3

Predictions from fitted regression models for SPR data. 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. Uncertainty interval is +/- 1.97 standard error.*

Term	5-gram	TXL	GRNN	GPT-2
Intercept	344.1 [329.2, 359]	345.9 [331, 360.8]	345.9 [331.1, 360.7]	345.9 [331, 360.8]
Surprisal	1.1 [0.5, 1.8]	0.8 [0.2, 1.4]	1.2 [0.6, 1.8]	0.8 [0.2, 1.4]
Length	1.9 [0.4, 3.4]	1.9 [0.5, 3.4]	1.8 [0.4, 3.3]	1.9 [0.5, 3.4]
Frequency	0.8 [-0.2, 1.8]	0.4 [-0.5, 1.2]	0.6 [-0.2, 1.4]	0.4 [-0.5, 1.2]
Surp x Length	-0.2 [-0.5, 0.1]	-0.1 [-0.4, 0.2]	-0.3 [-0.6, 0]	-0.1 [-0.4, 0.2]
Freq x Length	-0.3 [-0.8, 0.1]	-0.2 [-0.6, 0.1]	-0.4 [-0.7, 0]	-0.2 [-0.6, 0.1]
Past Surprisal	0.6 [-0.1, 1.2]	0.5 [0, 1.1]	0.6 [0, 1.2]	0.5 [0, 1.1]
Past Length	0.8 [-0.7, 2.3]	0.8 [-0.7, 2.3]	0.7 [-0.8, 2.2]	0.8 [-0.7, 2.3]
Past Freq	0.8 [-0.2, 1.7]	0.6 [-0.3, 1.4]	0.6 [-0.2, 1.4]	0.6 [-0.3, 1.4]
Past Surp x Length	-0.1 [-0.5, 0.2]	0.1 [-0.2, 0.3]	-0.1 [-0.4, 0.2]	0.1 [-0.2, 0.3]
Past Freq x Length	-0.1 [-0.6, 0.3]	0.1 [-0.3, 0.5]	-0.1 [-0.4, 0.3]	0.1 [-0.3, 0.5]
2Past Surprisal	0.4 [-0.3, 1]	-0.3 [-0.9, 0.2]	-0.2 [-0.8, 0.4]	-0.3 [-0.9, 0.2]
2Past Length	1.8 [0.4, 3.2]	1.8 [0.4, 3.2]	1.8 [0.4, 3.2]	1.8 [0.4, 3.2]
2Past Freq	0.8 [-0.2, 1.8]	0.1 [-0.7, 0.9]	0.2 [-0.6, 1]	0.1 [-0.7, 0.9]
2Past Surp x Length	-0.1 [-0.4, 0.2]	-0.1 [-0.4, 0.2]	-0.2 [-0.5, 0.1]	-0.1 [-0.4, 0.2]
2Past Freq x Length	-0.1 [-0.6, 0.3]	-0.1 [-0.5, 0.2]	-0.2 [-0.5, 0.2]	-0.1 [-0.5, 0.2]
3Past Surprisal	-0.2 [-0.9, 0.5]	-0.6 [-1.1, 0]	-0.7 [-1.2, -0.1]	-0.6 [-1.1, 0]
3Past Length	1.3 [-0.1, 2.7]	1.2 [-0.2, 2.6]	1.3 [0, 2.7]	1.2 [-0.2, 2.6]
3Past Freq	0.4 [-0.6, 1.4]	0.2 [-0.6, 1]	0.2 [-0.6, 1]	0.2 [-0.6, 1]
3Past Surp x Length	-0.2 [-0.5, 0.2]	0 [-0.3, 0.2]	-0.1 [-0.4, 0.2]	0 [-0.3, 0.2]
3Past Freq x Length	-0.1 [-0.6, 0.4]	0.1 [-0.3, 0.4]	0 [-0.4, 0.4]	0.1 [-0.3, 0.4]

read all the words because they can correct their mistakes and move on. We tested this method on the Natural Stories corpus, showing that, despite the oddness of the task, participants can read and understand a 1000 word story in this method. We additionally showed that the RTs generated in the Maze task show a linear relationship between the RT of a word and it's surprisal, but no relationship with the surprisal of the previous word. This provides additional evidence for the argument that the Maze task forces very incremental processing (Forster et al., 2009).

Actually the potential applicability and value is much broader, for any question where nailing down the locus on incremental processing difficulty is important, little spillover is very valuable.

This extreme incrementality makes the Maze task a good target for any question that requires precisely determining the locus of incremental processing difficulty and thus benefits from the lack of spillover. It shows broadly the same effects as other methods; however the differences in scale of suprisal effects and the lack of frequency effects detected here are a

Table 4

Results of model comparisons on SPR data. Each row shows the additional predictive value gained from adding that model to another model. F values and p values from anova tests are reported. We also report log likelihoods of models with only one surprisal source.

Model	over 5-gram	over GRNN	over TXL	over GPT-2	Log Lik	r_squared
5-gram		6 (p<0.001)	8 (p<0.001)	4 (p=0.003)	-47899	0.01
GRNN	4 (p=0.007)		3 (p=0.031)	3 (p=0.028)	-47904	0.01
TXL	3 (p=0.023)	1 (p=0.697)		4 (p=0.007)	-47909	0.01
GPT-2	7 (p<0.001)	8 (p<0.001)	11 (p<0.001)		-47894	0.02

reason to explore more. Comparisons between different processing on the same materials could be useful for identifying how task demands influence language processing (ex. Bartek et al., 2011).

While the error-correction paradigm is crucial to running long materials, it also provides some benefits even on shorter materials when using A-maze with variable participant populations. The error-correction compensates for some of the shortcomings of A-maze identified in Boyce et al. (2020); poor distractors might still cause participants to make unavoidable mistakes, but they will still see the rest of the sentence, which may reduce frustration. One questions that remains is whether this post-mistake data is useable, which depends on whether RTs from a few words after a mistake show any differences from RTs before any mistakes. Whether post-mistake data is high-quality and trustworthy enough to be included in analyses is a hard-to-assess question of potential interest.

Even if post-mistake data is not analysed, it can be used to distinguish between errors due to inattentive participants or merely specific bad distractors early in the sentence. Researchers can calculate a per-word error rate for each participant; high per-word error rates are consistent with guessing, low error rates with errors clustered early the sentence are consistent with poor distractors. This per-word error rate metric measures participants task accuracy and provides a convenient and clear-cut way of controlling data quality after the fact.

Error-correction Maze also starts to reduce perverse incentives from the desire to complete the task quickly. With traditional Maze, clicking randomly will likely lead to a mistake, which will cause a participant to skip ahead to the next sentence. With error-correction Maze, randomly jamming buttons takes more effort, but is still faster than doing the task. In discussing this work, we received 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.

Between the distractor auto-generation process introduced in Boyce et al. (2020) and the error-correction paradigm introduced here, the Maze paradigm is now easy to use on a wide range of materials. The ease of use, large effects, and forced incrementality make Maze a good complement to existing incremental processing methods. We encourage researchers to consider Maze as an option for doing incremental processing work.

References

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Human Perception & Performance*, *37*(5), 1178.

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.

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

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the r package brms. *The R Journal*, *10*(1), 395–411.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv:1901.02860 [Cs, Stat].* Retrieved from http://arxiv.org/abs/1901.02860

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

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

Freedman, S. E., & Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition*, *19*(2), 101–131.

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.*

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 70–76). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.10

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* (pp. 10–18). Salt Lake City, Utah: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0102

Grodner, D., & Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science*, *29*(2), 261–290.

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).

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. *arXiv:2005.03692 [Cs]*. Retrieved from http://arxiv.org/abs/2005.03692

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

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284.

Koornneef, A. W., & van Berkum, J. J. A. (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. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results, 11.

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

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of*

Memory and Language, *32*, 692–715.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). Language Models are Unsupervised Multitask Learners, 24.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The Effects of Frequency and Predictability on Eye Fixations in Reading: Implications for the E-Z Reader Model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 720–732.

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/

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4086–4094). Minneapolis, Minnesota: Association for Computational Linguistics.

Sloggett, S., Handel, N. V., & Rysling, A. (2020). A-maze by any other name. In *CUNY*.

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

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*, 71–86.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*, 69–90.

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

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv:2006.01912 [Cs]*. Retrieved from http://arxiv.org/abs/2006.01912

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.

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.

Appendix

The beginning of one of the stories. This is the first 200 words of a 1000 word story.

Tulip mania was a period in the Dutch Golden Age during which contract prices for bulbs of the recently introduced tulip reached extraordinarily high levels and then suddenly collapsed. At the peak of tulip mania in February sixteen thirty-seven, tulip contracts sold for more than ten times the annual income of a skilled craftsman. It is generally considered the first recorded economic bubble. The tulip, introduced to Europe in the mid sixteenth century from the Ottoman Empire, became very popular in the United Provinces, which we now know as the Netherlands. Tulip cultivation in the United Provinces is generally thought to have started in earnest around fifteen ninety-three, after the Flemish botanist Charles de l'Ecluse had taken up a post at the University of Leiden and established a botanical garden, which is famous as one of the oldest in the world. There, he planted his collection of tulip bulbs that the Emperor's ambassador sent to him from Turkey, which were able to tolerate the harsher conditions of the northern climate. It was shortly thereafter that the tulips began to grow in popularity. The flower rapidly became a coveted luxury item and a status symbol, and a profusion of varieties followed.

The first 2 out of the 6 comprehension questions. When did tulip mania reach its peak? 1630's, 1730's From which country did tulips come to Europe? Turkey, Egypt