A-maze of Natural Stories: Comprehension and surprisal in the Maze task TODO title

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

TODO needs work 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.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.

Keywords: TODO
Word count: TODO

A-maze of Natural Stories: Comprehension and surprisal in the Maze task TODO title

Introduction

TODO overall: go through Roger's comments! TODO fix tense!

Sometimes we stumble in reading, where we expected a sentence to go in one direction and it did not, and then are at a parsing dead-end where we have to read a sentence again. The words that cause stumbles are the outliers on a distribution of words difficulty to read, within fluent reading, there is variation in how easy words are to read and process in their context. We're able to read many long or surprising words without noticing a problem, but these less expected words do take longer to process as we rebuild our burgeoning mental model of the sentence. Fortunately for fluent readers and unfortunately for studying language, this process of adjustment is very quick, with millisecond scale differences in reading time between words.

We don't have a thorough understanding of how the mind processes language, but we can learn something about what is going on from measuring how long it takes in different circumstances. This relies on a linking hypothesis that processing time reflects something about how much processing is going on, even if we don't know whether this is character recognition, memory retrieval, or parsing. This is enough to compare theories that predict the relative difficulty, and thus relative processing time, of two words in different contexts.

For instance, different theories about what makes object relative clauses harder to understand than subject relative clauses make different predictions about which words are the loci of the overall difficulty and slower reading times associated with object relatives (Grodner & Gibson, 2005; Staub, 2010; Traxler, Morris, & Seely, 2002). Measures of reading time can also inform theories about the time course of processing (i.e. which steps are parallel versus serial, Bartek, Lewis, Vasishth, and Smith (2011)) or the functional form of relationships between word characteristics and processing time (Smith & Levy, 2013).

Some of these theories rely on being able to attribute processing slowdowns to a particular word. Determining that object relatives are overall slower that subject relatives is easy, even an imprecise measure of reading time will determine that the same set of words in a different order took longer to read on a sentence level. However, many language processing theories make specific (and contrasting) theories about which words in a sentence should be harder to process. To adjudicate these theories, we want methods that are *localized*, where it is easy to determine which word is responsible for an observed slow-down in reading time. Ideally, a longer reading time on a word would be an indication of that word's increased difficulty, and not the lingering signal of a prior word's increased difficulty. When the signal isn't localized, advanced analysis techniques may be required to disentangle the slow-downs (Shain & Schuler, 2018).

Incremental processing methods

Behavioral methods that measure how long it takes to read each word in a sentence are referred to as incremental processing methods, and their dependent measures are called

reading times or reaction times, both abbreviated RT. The two most commonly used methods are eye-tracking and self-paced reading, but both of these suffer from a lack of localization. In eye-tracking, participants read a text on a screen naturally, while their eye-movements are recorded (Rayner, 1998). The downside of this natural reading is that people often skip short words and look ahead or look back as they read, the dynamics of which make it hard to isolate effects (Frazier & Rayner, 1982; Levy, Bicknell, Slattery, & Rayner, 2009; Rayner, Ashby, Pollatsek, & Reichle, 2004). Self-paced reading (SPR) is a more controlled process where a participant sees one word on screen at a time and presses a button to see the next word instead. Even in this method, 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). Both these methods are prone to spillover, as the time to plan motor movements like saccades or button presses is substantial compared to the time it takes to recognize a word, these movements are often initiated before a word is fully processed, so by the time any late processing difficulties occur, the reader may already be several words along.

the Maze task. An alternative method that is designed to increase localization at the expense of naturalness is the Maze task (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 time to selection is treated as the reaction time (RT). Forster et al. (2009) introduces two versions of the Maze task: lexical L-maze where the distractors are nonce words and grammatical G-maze where the distractors are real words that don't fit with the context of the sentence so far. Theoretically, participants must fully integrate each word into the sentence in order to confidently select it; this may require mentally reparsing previous material in order to allow the integration and selection of a disambiguating word. Forster et al. (2009) call this forced incremental processing to distinguish from other incremental processing methods where words can be passively read before later committing to a parse. This idea of strong localization 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 has not been 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, and thus likely to be judged infelicitous by human readers. 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.

Another recent variant of the Maze task is interpolated or I-maze, which uses a mix of real word distractors (generated via the A-maze process) and nonce word distractors (Vani, Wilcox, & Levy, 2021). The presence of real word distractors encourages close attention to the sentential context, while nonce words can be used as distractors where the word in the

sentence is itself ungrammatical or highly unexpected. Vani et al. (2021) used I-maze to compare object- and subject- relative clauses and Wilcox, Vani, and Levy (2021) used it to compare differences in RT for pairs of grammatical and ungrammatical sentences NLP language models predictions.

Frequency, length, and surprisal effects

One way of assessing how much a method localized RT effects is to look at how strongly properties of a word correlate with the RT of that word compared with RTs of downstream words later in the sentence. A couple of properties known to influence reading time are a word's length and overall frequency in a language, as longer and less common words take longer to process (Kliegl, Grabner, Rolfs, & Engbert, 2004). Another measure is predictability, or how expected a word is given the context, with more predictable words being read faster (Rayner et al., 2004). A word can have low predictability for a number of reasons: it could be low frequency, semantically unexpected, the start of a low-frequency syntactic construction, or a word that disambiguates prior words to the less common parse. Many targeted effects of interest are essentially looking at specific features that contribute to how predictable or unpredictable a word is. This means that how good a method is at detecting and localizing effects of predictability is a key aspect of how useful it is as measure of incremental processing. With al these factors potentially contributing to predictability, how is predictability as a whole measured? The typical method is to use language models that are trained on large corpora of language to predict what word comes next in a sentence. A variety of pre-trained models exist, that vary in their internals, but all of them will generate assessments of how predictable words are. Predictability is often represented in terms of bits of surprisal, which is the negative log probability of a word (1 bit of surprisal means a word is expected to occur half the time, 2 bits is 1/4 of the time etc). The functional form of the relationship between RTs from eye-tracking and SPR corpora and the predictability of the words is linear in terms of surprisal (Goodkind & Bicknell, 2018; Luke & Christianson, 2016; Smith & Levy, 2013; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). Due to spillover effects, this linear relationship also holds between the surprisal of a previous word and the RT on the current word.

Current experiment

The Maze task has thus far only been used on constructed sentences focusing on targeted effects and not on the sorts of long naturalistic passages used to assess the relationship between RT and surprisal. We want to test whether participants can read and understand stories while doing the demanding Maze task and also whether the RT profiles from the Maze task are similar to those from other methods.

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 are a severe test of A-maze, as these features may make it harder to auto-generate appropriate distractors and require focus from participants. If participants can succeed at the Maze task on Natural

Stories, 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.

We tweak the A-maze task to accomodate these longer passages and then have participants read the passages in the Maze. We compare our A-maze results with SPR data collected on this corpus by Futrell et al. (2020). We find that participants were able to read and understand these long passages using A-maze, and their RT profiles show a similar pattern to SPR but with less noise and spillover

Error-correction maze

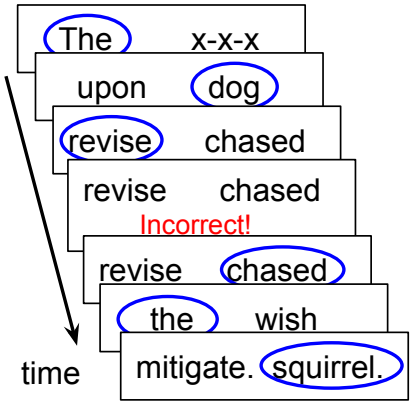

Figure 1. Schematic of error-correction Maze. A participant reads a sentence word by word, choosing the correct word at each time point (selections marked in blue ovals). When they make a mistake, an error message is displayed, so they try again and continue with the sentence.

One of the benefits of the Maze task is that it forces incremental processing by having participants make an active choice about what the next word is. But what if they choose incorrectly? In the traditional Maze paradigm, there isn't a way to handle this; a mistake indicates they're not understanding the sentence properly, so the sentence ends, and the participant moves on to the next item (Forster et al., 2009). An advantage of this is that participants who contribute RT data are very likely to have understand the sentence up to that point (it's hard to choose correctly in G-maze otherwise). This contrasts with other methods, where determining whether participants are paying attention usually requires separate comprehension check questions, that are usually not used for Maze.

However, terminating sentences on errors means that 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 with long items. 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% of participants would complete each 15-word sentence, but only 61% would complete a 50 word vignette, and 13% would complete a 200 word passage. 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. Error-correction Maze is a change in experimental procedure, and is independent of what types of distractors are used. This error-correction presentation is used here with A-maze, but would also work with G-maze or I-maze.

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 used the texts from the Natural Stories corpus (Futrell et al., 2020) and generated A-maze distractors for them. 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.

To generate distractors, we first split the corpora up into sentences, and then ran the sentences through the A-maze generation process. We used an updated version of the codebase from Boyce et al. (2020). This newer version had the capability to match the greater variety of punctuation present in this corpus (updated auto-generation code at https://github.com/vboyce/Maze). We took the auto-generated distractors as they were, without checking them for quality.

Participants

We recruited 100 participants from Amazon Mechanical Turk in April 2020, 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, and were debriefed, before getting 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

TODO top level description

In order to measure the relationship between a word's properties and it's RT, we fit models using surprisal, frequency, and length as predictors of RT. We considered these predictors from both the current and past word to look for the possibility of spill over effects. We fit generalized additive models (GAMs) to look at the functional form of the surprisal-RT relationship and linear models (LMs) to look at the size of the effects. As described in the Results section, we base all models only on data from participants who choose correctly on at least 80% of the words in the Maze task.

TODO FIX ME 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).

Predictors. We created a set of predictor variables of frequency, word length, and surprisals from 4 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 rescaled it to be the log2 frequency of expected occurrences 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). These models cover a range of common architectures. We can infer robustness if models provide similar results, but also compare between models to look at how well each model fits human data. For all of these predictors, we consider both the predictor at the current word as well as lagged predictors from the previous word.

Exclusions. We exclude the first word of every sentence because it had an x-x-x distractor, which leaves 9782 words. We exclude words for which we don't have surprisal or frequency information, leaving 8489 words. We additionally exclude words that any model treated as being composed of multiple tokens. While surprisals should be additive, the summing the surprisals over these tokens gives some unreasonable responses. 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 any model treated as multi-token, leaving 7512 words. This primarily excludes words with punctuation. (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.)

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.

Model specification. 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. TODO edit for bootstrapping

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. TODO cite LM

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.

SPR-Maze correlation

We additionally compare how correlated the Maze and SPR results are to each other, in comparison to within-Maze and within-SPR correlations. For Maze, within each story, we randomly split subjects into two halves. Within each half, we calculate a per-word average for each word and then a per-sentence average RT across word averages. We calculate a within-Maze correlation between these two halves. To avoid differences due to dataset size, we downsample the SPR data choosing a number of participants equal to the number we have for Maze. We then use the same procedure on this subset as for Maze to get a within-SPR correlation. For between Maze-SPR correlation, we take the average correlation across each of the 4 pairs of Maze half and SPR half.

Results

Reading stories in the Maze task

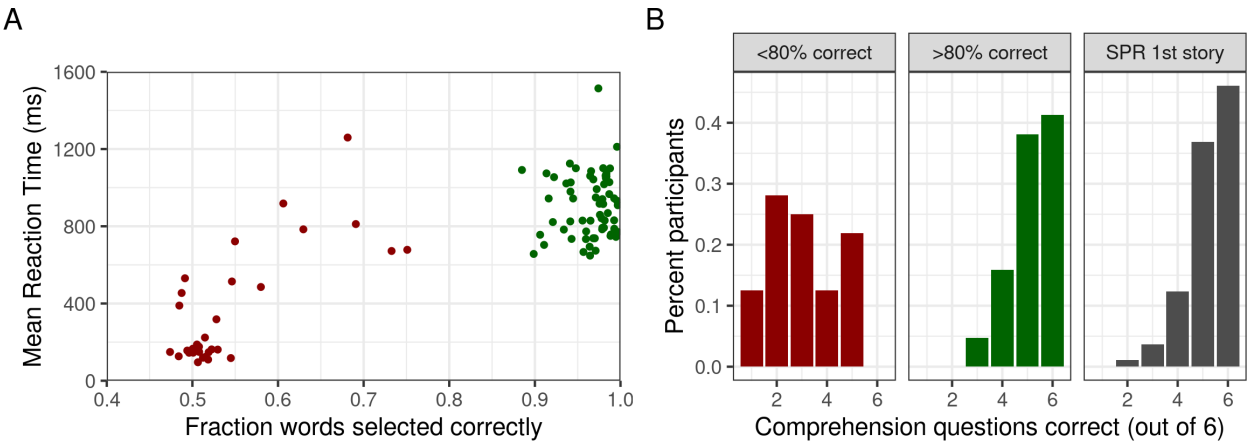

Figure 2. A. Correlation between a participant's accuracy on the Maze task (fraction of words selected correctly) and their average reaction time (in ms). Many participants (marked in green) chose the correct word >80% of the time; others (in red) appear to be randomly guessing and were excluded from 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.

We first looked at participants overall performance on the Maze task and comrehension questions, in order to see if participants were able to read and understand a story in the

Maze task. Many participants were able to complete the Maze task with a high degree of accuracy and answer the comprehension questions correctly.

Participant accuracy, or how often participants choose the correct word over the distractor, reflects both the quality of the distractors and the focus and skill of the participant. 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; this excluded 260 words, or 0.27% of trials.) As seen in Figure 2A, one cluster of participants (marked in green) made 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 in eye-tracking or SPR. 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 who 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 for A-maze 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

TODO: question: Does the current order of all Maze then all SPR results make sense? Or would be it better to have Maze-GAM, SPR-GAM, Maze-LM, SPR-LM, etc?

Given that the Maze task worked, we next consider the relationship between surprisals and Maze RTs. Surprisal, a measure of overall word predictability in context, is linearly related to RT in eye-tracking and SPR. If Maze is measuring the same language processes,

we would expect to see a linear relationship in the Maze task. Additionally, surprisal can be used to assess spillover, but comparing the predictive value of current and past word surprisal on current word reading time.

To assess the shape of the RT-surprisal relationship, we fit generalized additive models (GAMs). 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. We use surprisals generated by 4 different language models for robustness. (See Methods for details on language models, exclusions, and model fit.)

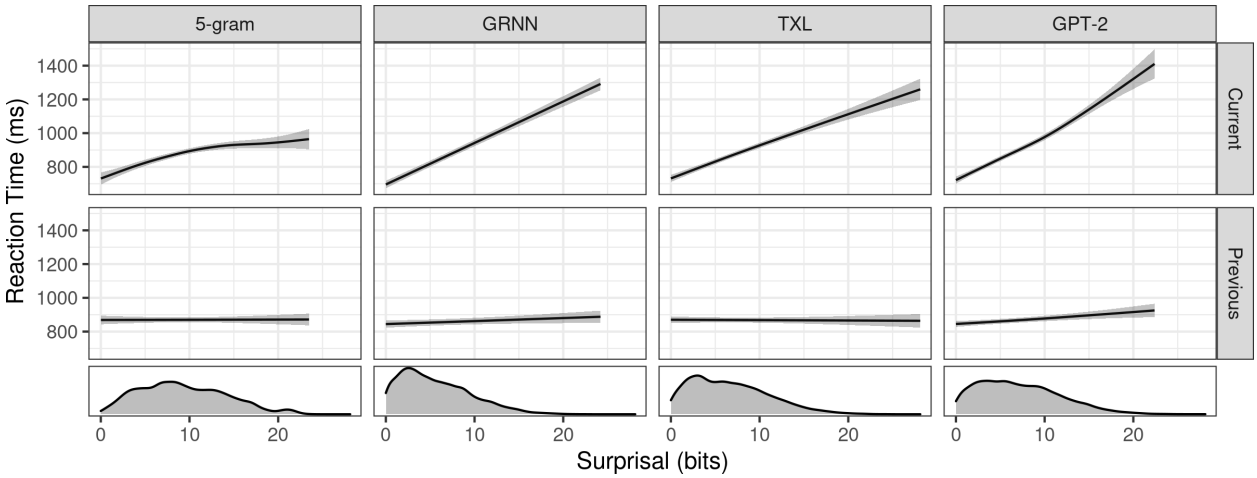

Figure 3. GAM predictions of reaction time (RT) as a function of either current word surprisal (top) or previous word surprisal (bottom). Density of data is shown along the x-axis. For each of the 4 language models used, there is a linear relationship between current word surprisal and RT (at least when there is enough data). The relationship between previous word surprisal and RT is much flatter.

The GAMs predicted relationship between current and previous word surprisals are shown in Figure 3; this is the main effect of either current or previous surprisal for an average length and frequency word, if the other surprisal value is 0 (TODO Not sure this is the correct interpretation! – https://rdrr.io/cran/tidymv/man/get_gam_predictions.html is the function used for prediction). 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 flatter relationship between previous word surprisal and RT. This is a sign of localization as the previous word's surprisal is not affecting RT much. The linear relationship matches that found in SPR and eye-tracking. TODO comment on freq, length effects / non-linearities!

Given that the GAM models show a roughly linear relationship between RT and surprisal, we are justified in using surprisal as a predictor in a linear model to quantify its effect. In addition to surprisal, we also include the commonly studied effects of frequency and length as well as surprisal x length and frequency x length interactions. For all of these, we include the predictors for the current and previous word, and we centered, but did not

rescale, all predictors. (See Methods for more details on these predictors and model fit process.)

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. Interval is 2.5th quantile to 97.5th quantile of model draws.

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 shown in Table 1, we found large effects of surprisal and length, but minimal effects of frequency. 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, an indication that spillover is limited and effects are strongly localized.

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. 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. Log likelihood, which is a measure of fit to data, shows a similar hierarchy, as GPT-2 is better than GRNN is better than TXL is better than 5-gram.

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 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.

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 between 1-surprisal-source and 2-source models 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

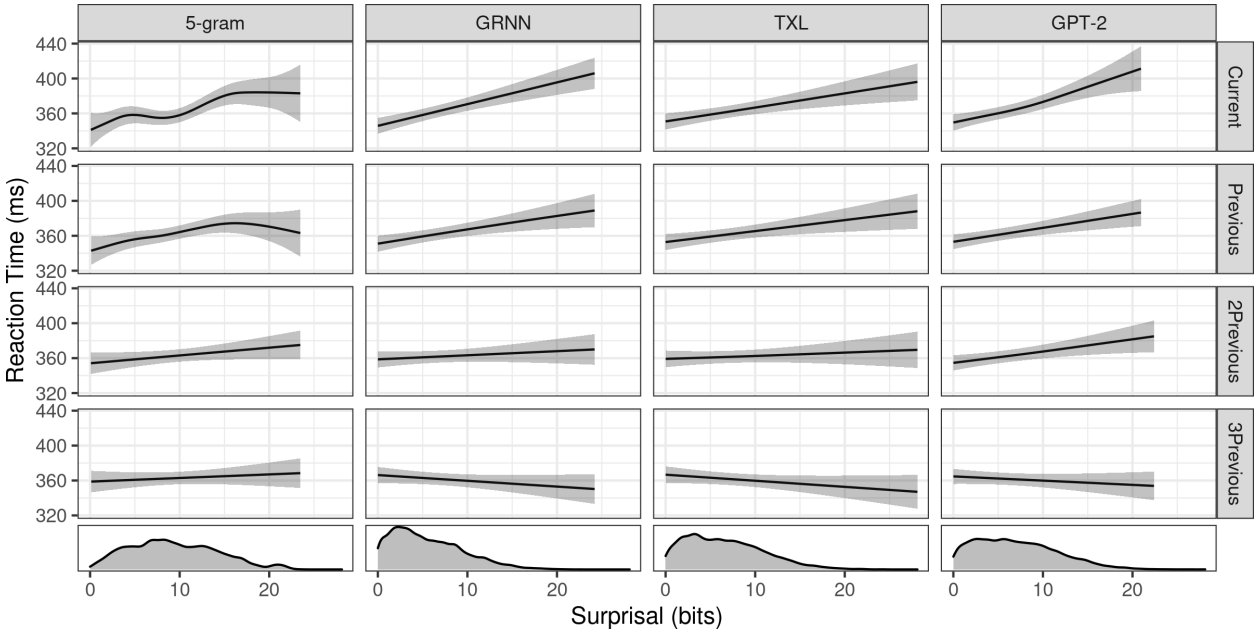

Figure 4. GAM predictions of reaction time (RT) for SPR data from Futrell et al. (2020) as a function of current word surprisal (top) or the surprisal of an earlier word, up to 3 words back. Density of data is shown along the x-axis.

Similarly to the model comparison for Maze, we also conducted a model comparison on the SPR data shown in Table 4. 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. 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. 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

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	361.6 [344.5, 378.6]	363.9 [346.9, 380.9]	363.8 [346.8, 380.8]	363.9 [346.9, 380
Surprisal	1.1 [0.2, 2.1]	1.1 [0.3, 1.9]	1.8 [1, 2.7]	1.1 [0.3, 1.9]
Length	2.1 [0, 4.2]	2.2 [0.1, 4.2]	2 [-0.1, 4]	2.2 [0.1, 4.2]
Frequency	1.4 [0, 2.8]	1.2 [0.1, 2.4]	1.6 [0.5, 2.8]	1.2 [0.1, 2.4]
Surp x Length	-0.2 [-0.6, 0.3]	0.1 [-0.3, 0.4]	-0.2 [-0.6, 0.1]	0.1 [-0.3, 0.4]
Freq x Length	-0.4 [-1, 0.2]	-0.2 [-0.7, 0.3]	-0.4 [-0.9, 0.1]	-0.2 [-0.7, 0.3]
Past Surprisal	1 [0.1, 1.9]	0.7 [0, 1.5]	0.9 [0.1, 1.7]	0.7 [0, 1.5]
Past Length	0.1 [-2, 2.1]	0.1 [-1.9, 2.1]	-0.1 [-2.1, 1.9]	0.1 [-1.9, 2.1]
Past Freq	1.5 [0.2, 2.9]	1.1 [0, 2.2]	1.1 [0, 2.2]	1.1 [0, 2.2]
Past Surp x Length	-0.1 [-0.5, 0.3]	0.2 [-0.2, 0.5]	0 [-0.4, 0.3]	0.2 [-0.2, 0.5]
Past Freq x Length	-0.2 [-0.8, 0.5]	0.1 [-0.4, 0.6]	-0.1 [-0.6, 0.4]	0.1 [-0.4, 0.6]
2Past Surprisal	0.6 [-0.4, 1.5]	-0.2 [-1, 0.6]	0 [-0.8, 0.8]	-0.2 [-1, 0.6]
2Past Length	2.2 [0.3, 4.2]	2.1 [0.2, 4]	2.1 [0.2, 4]	2.1 [0.2, 4]
2Past Freq	1.5 [0.2, 2.8]	0.7 [-0.5, 1.8]	0.8 [-0.3, 1.9]	0.7 [-0.5, 1.8]
2Past Surp x Length	-0.3 [-0.7, 0.2]	0 [-0.3, 0.4]	-0.3 [-0.6, 0.1]	0 [-0.3, 0.4]
2Past Freq x Length	-0.3 [-1, 0.3]	0 [-0.5, 0.4]	-0.3 [-0.7, 0.2]	0 [-0.5, 0.4]
3Past Surprisal	-0.3 [-1.3, 0.6]	-0.9 [-1.7, -0.2]	-1 [-1.8, -0.2]	-0.9 [-1.7, -0.2]
3Past Length	1.1 [-0.9, 3]	0.8 [-1.1, 2.7]	1.1 [-0.9, 3]	0.8 [-1.1, 2.7]
3Past Freq	0.4 [-1, 1.7]	0 [-1.2, 1.1]	0 [-1.1, 1.1]	0 [-1.2, 1.1]
3Past Surp x Length	-0.5 [-0.9, 0]	-0.1 [-0.4, 0.3]	-0.3 [-0.6, 0.1]	-0.1 [-0.4, 0.3]
3Past Freq x Length	-0.4 [-1.1, 0.2]	-0.1 [-0.6, 0.4]	-0.2 [-0.7, 0.3]	-0.1 [-0.6, 0.4]

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 0.007 to 0.011. 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.

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		3 (p=0.032)	4 (p=0.001)	3 (p=0.033)	-51798	0.007
GRNN	7 (p<0.001)		6 (p<0.001)	2 (p=0.153)	-51790	0.009
TXL	3 (p=0.010)	0 (p=0.910)		1 (p=0.462)	-51801	0.007
GPT-2	10 (p<0.001)	5 (p<0.001)	10 (p<0.001)		-51783	0.011

An additional comparison between SPR and Maze is how correlated their reading times

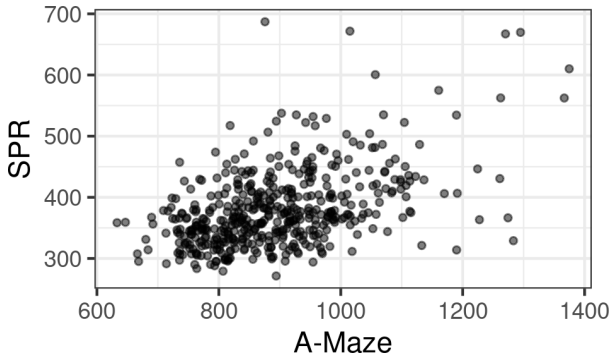

Figure 5. Correlation between SPR and Maze data. RTs were averaged across participants per word and then averaged together within each sentence. RTs in ms.

are; that is, do the methods pick up on the same effects. If so, we would expect that the sentences that take longer to read in one method also take longer in the other. We calculate the average RT at the sentence level (see Methods for details). The correlation between Maze and SPR is 0.26, compared to 0.23 within SPR and 0.37 within Maze. See Figure 5 for a visual comparison of overall Maze versus SPR RTs.

Discussion

We introduce error-correction Maze, a tweak on the presentation of Maze materials that makes Maze feasible for multi-sentence passages. We used A-maze distractors and this error-correction presentation to gather data on participants reading the Natural Stories in the Maze. We found that participants could read and comprehend the 1000 word stories, despite the slowness and added overhead of reading in the Maze task. This expands the domain of materials used with Maze beyond targeted single-sentence items to longer, naturalistic texts with sentence-to-sentence coherency.

We took advantage of the pre-existing SPR corpus on Natural Stories to compare the RT profiles between Maze and SPR. While these are not completely comparably as the subject populations differed, we are able to get a sense of the relationship between SPR and Maze results. Maze and SPR pick up on similar features in words, as shown by the correlations between Maze and SPR, and the fact that surprisal and length are predictive of RTs in both. However, the effect sizes of RT and surprisal are much larger is Maze, and the effects are more strongly localized on a single word. Further comparisons between different processing on the same materials could be useful for identifying how task demands influence language processing (ex. Bartek et al., 2011).

We find good localization on Maze, but it's worth noting that it's not absolute. Some of the models show small but statistically reliable relationships with surprisal from a past word. On the whole, however, these results support the idea that Maze forces language processing to be close to word-by-word, and that it's a relatively safe assumption that the RT of a word primarily reflects its own properties and not those of earlier words.

This extreme incrementality makes the Maze task a good target for any question that

requires precisely determining the locus of incremental processing difficulty.

Remarks about error-correction maze

With that in mind, we have a few remarks on the benefits of error-correction Maze and some avenues for further work on this method. Error-correction Maze makes for a smoother process. It dampens the effect of rare poor A-maze distractors; a distractor that fits the context might still cause participants to make an unavoidable mistake, but the mistake is less disruptive, as they still see the rest of the sentence. An open question for researchers is whether data from after a participant makes a mistake is useable, that is, does it show the same profile as pre-errors words, or are there traces from recovering from the mistake? 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 included in analyses, having this complete information can distinguish between errors due to inattentive participants and errors from specific bad distractors. 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 error-correcting Maze and other innovations on the Maze paradigm such as A-maze and I-maze, the Maze task is versatile and can be used or adapted for a wide range of materials and questions of interest. We encourage researchers to use Maze as an incremental processing method, alone or in comparison with other methods.

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., 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.

Shain, C., & Schuler, W. (2018). Deconvolutional Time Series Regression: A Technique for Modeling Temporally Diffuse Effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2679–2689). Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1288

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.

Vani, P., Wilcox, E. G., & Levy, R. (2021). Using the Interpolated Maze Task to Assess Incremental Processing in English Relative Clauses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

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

Wilcox, E., Vani, P., & Levy, R. (2021). A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 939–952). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.76

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