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

Veronica Boyce (MIT/Stanford), Roger Levy (MIT)
vboyce@stanford.edu

A-maze is a new method for measuring incremental sentence processing that can localize slowdowns related to syntactic ambiguities (Boyce et al. 2020; Sloggett et al. 2020). We test A-maze on the Natural Stories corpus (Futrell et al. 2017) and find that people can comprehend what they read during the Maze task. Moreover, the Maze 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.

The Maze task (Forster et al. 2009; Witzel et al. 2012) is an incremental processing method where participants read a sentence word by word (Figure 1). For each word position, participants see two words, one of which is the next word in the sentence and one of which is a distractor. Participants press a key to indicate which word continues the sentence; the time between key presses (reaction time, or RT) is the dependent measure. Traditionally, when a participant makes a mistake, the sentence stops and they move on to the next item. In order to present coherent texts, we instead have participants correct their mistakes (Figure 1). When a participant makes a mistake, they see an error message and must press the correct key to continue with the sentence. This way, participants see all the content and can follow the story, allowing us to test multi-sentence materials.

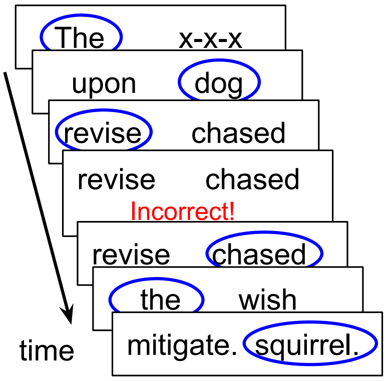
The Natural Stories Corpus consists of 10 passages ($\sim$1000 words each) with 6 comprehension questions per passage (Futrell et al. 2017). We used the A-maze framework from Boyce et al. (2020) and the language model from Gulordava et al. (2018) to generate distractor words for the texts. We recruited 100 participants from Amazon Mechanical Turk; each participant read one passage in the Maze task and answered the corresponding comprehension questions.

We exclude 5 participants who do not report being native English speakers. Participants clustered into those who appear to randomly press keys to get through the task quickly and those who do the task more slowly and accurately (Figure 2); we exclude participants with less than 80% accuracy from analyses. Task accuracy is correlated with performance on the comprehension questions ($r^2 = .47$); of the 63 participants who had at least 80% accuracy on the Maze task, 50 got 5 or 6 of the 6 comprehension questions correct (Figure 4). People can understand and remember what they read while doing the Maze task successfully.

We also investigated the relationship between a word's predictability and its RT. We fit Bayesian regression models using suprisal, frequency, length, and surprisal:length and frequency:length interactions at both the current and previous word as predictors of RT (see Table 1 for details). We use 3 models for surprisals: a Kneser-Ney smoothed 5-gram model, a recurrent neural network model (GRNN) from Gulordava et al. (2018), and a transformer model (TXL) from Dai et al. (2019). Across all models, we see large effects of current word surprisal and length, which noticeably exceed effects of previous word predictors (Table 1). We see minimal to no frequency effects (consistent with Shain 2019). In a nested model comparison, we found GRNN and TXL encoded complementary information, but adding 5-gram surprisals did not improve a regression that already had GRNN. We also fit GAMs to look at the shape of the relationship between surprisal and RT; we find a linear relationship between RT and current word surprisal and no relationship with previous word surprisal (Figure 3). The linear relationship matches known effects from eye-tracking and self-paced reading (Smith & Levy 2013), but the lack of spillover indicates a potential advantage of Maze over eye-tracking and self-paced reading.

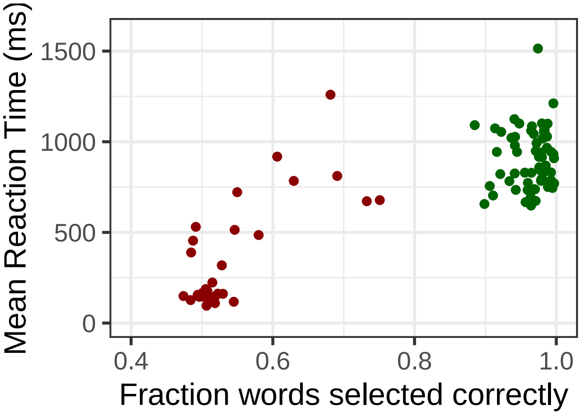
In summary, we provide an adaptation of the Maze task for long naturalistic texts, and show that participants can comprehend what they read during A-maze. Additionally, we find robust, localized surprisal effects, supporting Forster et al.'s (2009) argument that Maze forces highly incremental processing. Overall, this suggests that A-maze is a versatile alternative to eye-tracking and self-paced reading. Code and data at github.com/vboyce/amaze-natural-stories.

Figure 1: Maze with Error Correction



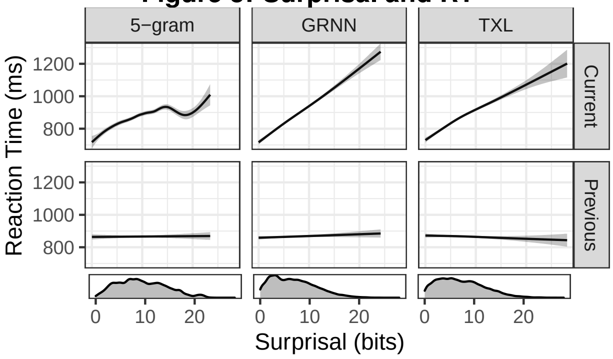
Participants see two words at a time and try to select the correct word. When the participant makes a mistake, they must correct it to continue. Blue circles indicate selected words.

Figure 2: Accuracy versus RT



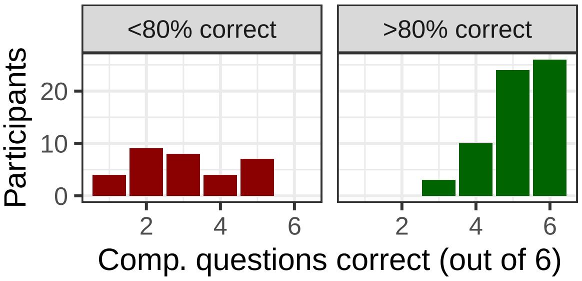
Correlation between participant's accuracy on the Maze task and RT. Participants with less than 80% accuracy (in red) were excluded from analyses.

Figure 3: Surprisal and RT



Current word surprisal is linearly predictive of RT, but previous word surprisal is not predictive of RT.

Figure 4: Comprehension question accuracy



Participants with high accuracy on the Maze task also perform well on comprehension questions.

Table 1: Regression Coefficients

	5-gram			GRNN			TXL		
	Est	CI	p	Est	CI	p	Est	CI	p
Intercept	865.3	[829.9, 902.9]	0.00	871.1	[837.9, 905.3]	0.00	870.8	[832.5, 907.8]	0.00
Surprisal	11.7	[9.3, 14.1]	0.00	23.7	[21, 26.5]	0.00	18.5	[16.1, 21.1]	0.00
Frequency	-2.9	[-6.3, 0.5]	0.10	2.9	[-0.2, 6]	0.06	0.4	[-2.7, 3.5]	0.79
Length	20.5	[15.4, 25.6]	0.00	18.5	[13.3, 23.7]	0.00	21.4	[16.2, 26.6]	0.00
Surprisal:Length	-2.0	[-3, -1]	0.00	-1.8	[-2.7, -0.9]	0.00	-1.4	[-2.2, -0.6]	0.00
Freq:Length	-1.0	[-2.5, 0.4]	0.16	-0.1	[-1.2, 1]	0.82	0.2	[-0.9, 1.2]	0.76
Past Surprisal	1.6	[-0.5, 3.6]	0.14	2.7	[0.8, 4.5]	0.00	0.8	[-0.9, 2.5]	0.40
Past Freq	2.6	[-0.1, 5.4]	0.06	1.9	[-0.2, 4.2]	0.08	1.2	[-1.1, 3.6]	0.30
Past Length	-4.8	[-9, -0.1]	0.04	-6.6	[-10.9, -2.1]	0.00	-5.2	[-9.3, -0.7]	0.03
Past Surp:Length	-0.2	[-1.2, 0.8]	0.72	-0.9	[-1.7, -0.2]	0.01	-0.6	[-1.3, 0.2]	0.13
Past Freq:Length	-1.0	[-2.3, 0.3]	0.15	-1.8	[-2.9, -0.8]	0.00	-1.5	[-2.6, -0.5]	0.01

Point estimates, credible intervals, and p-value equivalents. Surprisal was measured in bits, frequency in log_2 occurrences per billion words, and length in characters. All predictors were centered, and only single token words were included. Models were fit in BRMS with formula *rt* $\sim$ *surprisal*length + freq*length + past_surp*past_length + past_freq*past_length + (surprisal*length + freq*length + past_surp*past_length + past_freq*past_length | participant)+(1|word_id)*. Results shown for regression on pre-error data only.

References: Boyce et al (2020). *J Mem. Lang.* • Dai et al (2019) arXiv:1901.02860 • Forster et al (2009). *Behav. Res. Methods* • Futrell et al (2017). arXiv:1708.05763 • Gulordava et al (2018). *NAACL-HLT 2018* • Shain (2019). *NAACL-HLT 2019* • Sloggett et al (2020). *CUNY 2020* • Smith & Levy (2013). *Cognition* • Witzel et al (2012). *J Psycholinguist Res.*