I(nterpolated) Maze: High-sensitivity measurement of ungrammatical input processing
Pranali Vani, Ethan Wilcox and Roger Levy

Summary: The Maze task (Forster et al., 2009), in which an experimental participant "navigates" through a text by successively choosing the contextually appropriate target next word over an inappropriate one, is web-deployable with better power and sensitivity for detecting incremental-processing RT effects than self-paced reading (Boyce et al., 2020). In G(rammatical)-Maze, the distractor is a contextually inappropriate word; in L(exical)-Maze, the distractor is a nonce word. G-Maze is the more sensitive of the two, but it has a limitation for sentence processing research: using it to study the processing of ungrammatical input is problematic, since neither the target nor the distractor would be contextually appropriate. Here we introduce Interpolated Maze (I-Maze) to address this limitation. I-Maze mixes G-Maze and L-Maze distractors, with L-Maze distractors for ungrammatical words. We assess the three Maze variants in two English experiments: Wh-Cleft Structures (which tests syntactic category featural match) and and Main Verb / Reduced Relative Clause (MVRR) Garden-Path sentences (which tests expectations for parses, rather than true ungrammaticality). We find G-Maze and I-Maze more powerful than L-Maze. We also discover a novel result for MVRR Sentences: a critical-region garden-path disambiguation effect of relative clause reduction not only for ambiguous participles (*brought*), but also a smaller effect for unambiguous participles (*given*). Interestingly, these patterns also appear in surprisals of the GPT–2 neural language model.

I-Maze: I-Maze items were created by interpolating G-Maze and L-Maze distractors. L-Maze distractors were used for the second word of the sentence, all words in critical regions, and ~35% of the remaining words, in groups of two, where the first word appeared in all conditions. L-maze distractor words outside of the critical region provided a baseline estimate for L-Maze distractor times. L-Maze distractors were produced with Wuggy (Keuleers & Brysbaert, 2010).

Experiments: The Wh-Cleft experiment consisted of four conditions (Ex. 1). Slowdowns were expected in the *mismatch* conditions, relative to *match* conditions. The MVRR Garden-paths (Ex. 2) crossed (i) reduction of relative clauses and (ii) ambiguity of RC verb. We expected slowdowns in the reduced RC or ambiguous verb, as well as an interaction between conditions. Each experiment, which was hosted on Ibex Farm, included thirty Wh-Cleft items, twenty MVRR Garden-paths items and twenty fillers. Thirty participants were recruited on Amazon M-Turk.

Results: The results for Wh-Clefts can be seen in Figure 1. We find a significant effect of matching for G and I-Maze ($p<0.001$) but not for L-Maze. We also found a main effect where NP continuations were read faster ($p<0.001$, L-Maze $p<0.01$), likely because the content verb (e.g. "fixed") sets up stronger expectations for an object than a light verb (e.g. "did") does for its verbal complement. The results for MVRR Garden-paths can be seen in Figure 2. We find main effects of reduction for all Maze variants ($p<0.001$), a main effect of ambiguity for G-Maze and I-Maze ($p<0.01$, $p<0.001$), and the expected interaction for G-Maze and I-Maze ($p<0.001$, $p<0.05$). The increased power of the Maze task reveals a surprising novel effect, which is that unambiguous but reduced RCs produce slower reading times in downstream critical regions than unreduced ambiguous RCs even though both strings are consistent with only one syntactic parse. This effect is significant for G-Maze and I-Maze ($p<0.05$). Noisy-chanel models could account for this behavior, with reduced RCs being sufficiently rare that the processor reserves significant probability for nearby high prior-probability parses. However, we show that the incremental surprial values (negative log probabilities) of a contemporary neural language model also capture this behavior (Figure 4), suggesting that predictive processing models which do not maintain an incremental stack of parses can capture these effects just as well.

I(nterpolated) Maze: High-sensitivity measurement of ungrammatical input processing
Pranali Vani, Ethan Wilcox and Roger Levy

Figure 1: Cleft Sentences. Error bars are 95% confidence intervals

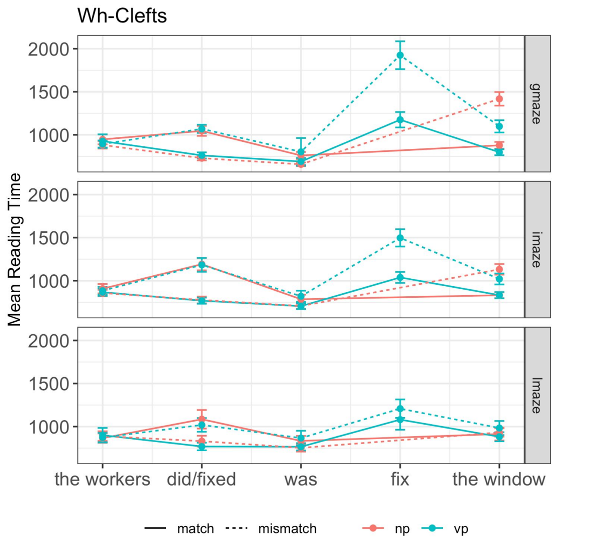
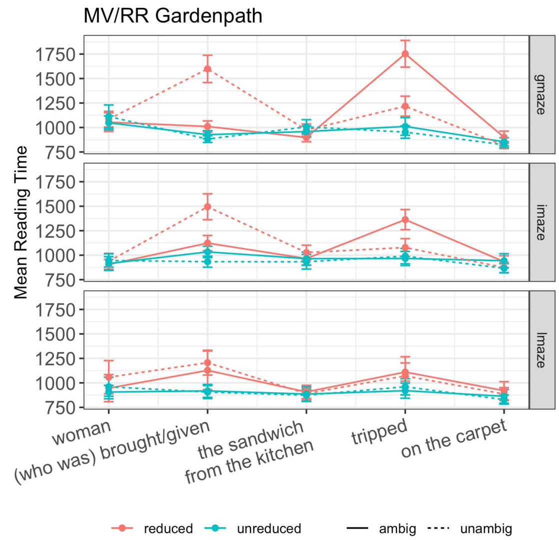


Figure 2: MV/RR Gardenpaths. Error bars are 95% confidence intervals



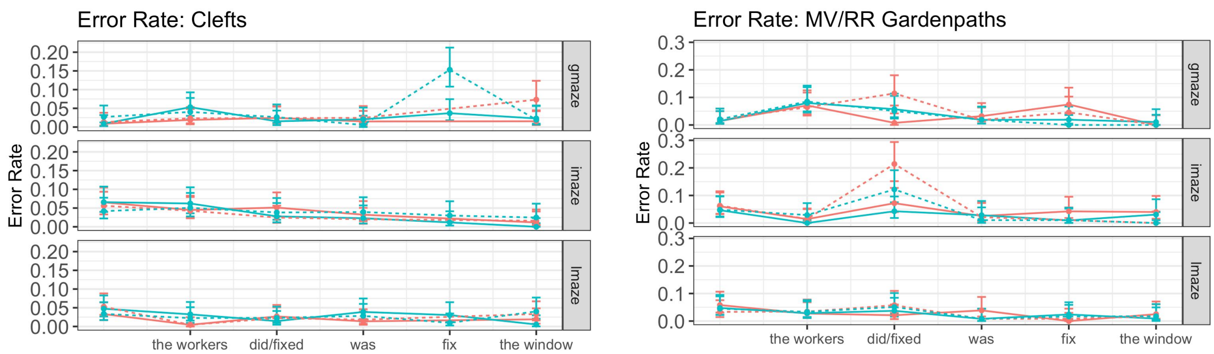
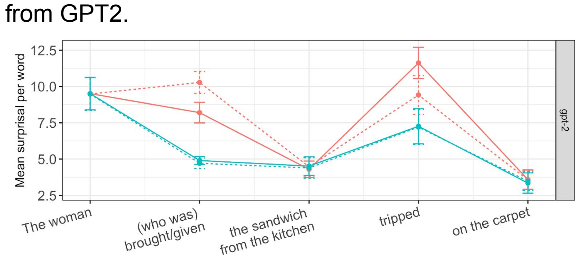


Figure 3: I-Maze and L-Maze critical region error rate in ungrammatical conditions is comparable to the other conditions. G-Maze critical region error rates are higher. This indicates that I-Maze eliminates the speed/accuracy tradeoff that G-maze is prone to, but produces results that are more sensitive to the different conditions than L-Maze.

Ex1: Wh-Cleft Sentences (critical regions are underlined)
What the workers did was <u>repair the window</u> [match/vp]
What the workers fixed was <u>repair the window</u> [mismatch/vp]
What the workers fixed was <u>the window</u> [match/np]
What the workers fixed was <u>repair the window</u> [mismatch/np]

Ex 2: MV/RR Gardenpath Sentences
The woman who was given the sandwich <u>tripped.</u> [unrecued / unambiguous]
The woman who was brough the sandwich <u>tripped</u>. [unreduced / unambiguous]
The woman given the sandwich <u>tripped</u>. [reduced/ unambiguous]
The woman brought the sandwich <u>tripped</u>. [reduced/ ambiguous]

Figure 4: Model results (MVRR Gardenpaths) Y-axis is mean surprisal values (-log(word|context)) derived from GPT2.



References: 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. • Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, *41*(1), 163-171. • Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*(3), 627-633.