Effects of foil processing, decision-making, and initial attention in the Maze task

John Duff[1], Pranav Anand[2], and Amanda Rysling[2] ([1] Lang. Sci. & Tech., Saarland; [2] Linguistics, UC Santa Cruz)

In [1]'s Maze task, at each word *w* of a stimulus, participants must decide between two proposed continuations shown side-by-side: a **target** (*w*) and a **foil** (a somehow-less-likely continuation). Maze decision latencies ("Maze RTs") on *w* closely track other reading time measures [1–4], without spill-over effects from previous words [4]. As a result of this precision, the Maze has become popular for crowd-sourced sentence-processing studies.

The RT on an accurate Maze trial could be modeled as the result of a consistent set of variable-latency processes: see Figure 1. However, this model has not yet been validated in full, and researchers have generally taken differences in Maze RTs across conditions to reflect differences in processing of the target (A–B). In this work, we demonstrate for the first time the additional contribution of foil processing (C–D) and decision-making (E).

Data comes from 2 previous experiments (Table 1) which used [3]'s A-Maze method to generate foils from high-surprisal continuations from [5]'s language model. We also calculated GPT-2 [7] surprisal (in bits) for every target and foil in context. Following e.g. [6] we assume that target processing durations (A–B) should be linearly related to target surprisal; we assume likewise that foil processing durations (C–D) should correlate with foil surprisal, and decision durations (E) should correlate with target–foil surprisal differences [9].

Procedure To isolate our effects of interest from known effects like target surprisal [3, 8], we first fit models with such effects (1a) over RTs on odd trials, then used these to calculate residual RTs on even trials. We then examined these residual RTs (1b), taking any remaining effects of foil surprisal or surprisal differences to reflect the influence of processes C–E.

However, if participants find the first word they attend to be suitable enough, they may choose it without examining the other, as in other forced choices [10]. Foil properties would then have less influence when the target is attended first. We thus also examine how foil surprisal interacts with trial diplay properties that may affect initial attention, like L–R ordering.

Results Residualizing models (Table 2) yielded credible effects of target surprisal, position within the stimulus, and trial number. Models over residuals provide evidence for interactions between foil surprisal effects and display properties (Table 3). Marginal estimates show that foil effects are additively larger when foils appear on the left, or opposite of the previous target (Table 4). (Participants, it seems, like to start at the left, or opposite their last choice.)

Figure 1 is thus only partially correct: Maze RTs do seem to come often from a chain of processes that includes foil processing and target–foil comparison, but they may also come from absolute judgments without comparison. We plan to validate these claims further with eye-tracking. For now, these results highlight the complex linguistic and decision processes that underlie Maze observations, and suggest that Maze experimenters should be careful to control for presentation order, foil surprisal, and surprisal differentials within critical regions.

Fig. 1: Hypothetical processes required for a Maze decision between target *w* and foil *w'* in context *c*.

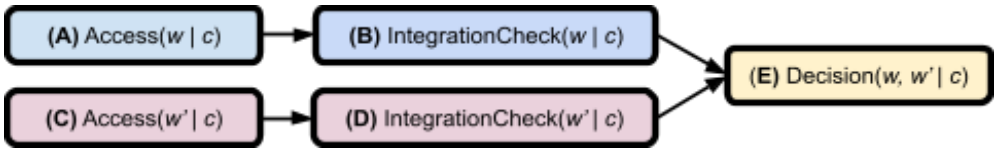

Table 1: Properties of the experimental data investigated here, taken from [11], Chapters 3 and 4.

Expt.	Participants	Distinct Items	Med. Stimulus Length (Range)	Total Correct RT Obs.
A	91	440	24 words (17–49)	180,309
B	143	336	21 words (5–26)	226,561

(1) Regression models fit in `brms` [12]

NB: All models used weakly-constrained priors, centered linear predictors, and dummy-coded binary predictors. We fit linear relationships to untransformed Maze decision times following [4].

a.
```
RT ~ Position * Trial + TargetSurp +
       (1 + Position * Trial + TargetSurp | Subject) + (1 | Item)
```

b.
```
Residual RT ~ (FoilSurp + SurpDiff) * TargetPosition * SideRepetition +
       (0 + (FoilSurp + SurpDiff) * TargetPosition * SideRepetition | Subject)
```

Expt	Parameter	$\hat{\beta}$	95% CrI
A	Intercept	790.89	(765.31, 816.17)
	TargetSurp	10.08	(9.42, 10.75)
	Position	2.27	(1.79, 2.76)
	Trial	-1.22	(-1.53, -0.92)
	Pos:Trial	0.02	(0.01, 0.04)
B	Intercept	826.24	(799.04, 853.21)
	TargetSurp	13.01	(12.12, 13.89)
	Position	4.86	(4.09, 5.62)
	Trial	-1.06	(-1.25, -0.87)
	Pos:Trial	0.02	(0.01, 0.03)

Table 2: Parameters in residualizing regressions.

Expt	Parameter	$\hat{\beta}$	95% CrI
A	FoilSurp:TargetPos	1.71	(0.74, 2.71)
	FoilSurp:SideRep	1.87	(0.73, 3.01)
	SurpDiff:TargetPos	-0.37	(-1.49, 0.73)
	SurpDiff:SideRep	-1.25	(-2.34, -0.16)
B	FoilSurp:TargetPos	2.44	(1.36, 3.52)
	FoilSurp:SideRep	2.83	(1.66, 4.01)
	SurpDiff:TargetPos	-1.35	(-2.51, -0.18)
	SurpDiff:SideRep	-2.56	(-3.89, -1.19)

Table 3: Notable interactions in critical regressions.

		FoilSurp		SurpDiff	
	Display Sequence	$\hat{\beta}$	95% CrI	$\hat{\beta}$	95% CrI
A	T F → T F	0.32	(-0.52, 1.16)	2.50	(1.26, 3.75)
	T F → F T	0.17	(-0.66, 1.00)	2.35	(1.24, 3.47)
	F T → F T	1.49	(0.35, 2.66)	3.67	(2.10, 5.27)
	F T → T F	-1.55	(-2.25, -0.86)	0.64	(-0.02, 1.29)
B	T F → T F	0.76	(-0.05, 1.59)	3.77	(2.22, 5.32)
	T F → F T	0.37	(-0.41, 1.17)	3.38	(1.93, 4.83)
	F T → F T	1.96	(0.98, 2.94)	4.97	(3.23, 6.65)
	F T → T F	-2.07	(-2.87, -1.26)	0.94	(0.07, 1.80)

Table 4: Marginal effects of foil surprisal parameters by target position sequence. (T F → F T means RTs from trials with a target on the right which were preceded by trials with a target on left.)

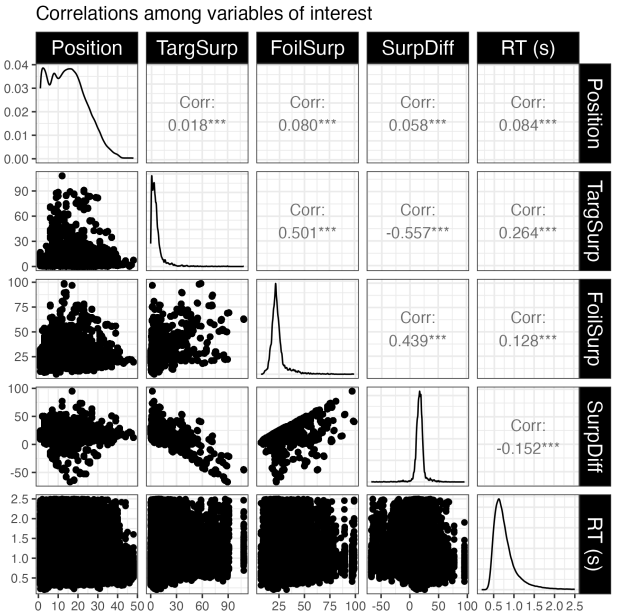

Fig. 2 (R): Distributions and joint distributions of parameters of interest in Expt. A.

References [1] Forster, Guerrera & Elliot (2009) *Behav Res Meth.* [2] Witzel, Witzel & Forster (2012) *J Psycholinguist Res.* [3] Boyce, Futrell & Levy (2020) *J Mem Lang.* [4] Boyce & Levy (2023) *Glossa Psycholing.* [5] Gulordava et al. (2018) *Proc NAACL-HLT*. [6] Shain et al. (2024) *PNAS.* [7] Radford et al. (2019) OpenAI. [8] Levinson (2023) *Proc ELM.* [9] Luce (1986) Oxford. [10] Starns, Chen & Staub (2017) *J Mem Lang*. [11] Duff (2023) UC Santa Cruz diss. [12] Bürkner (2017) *J Stat Soft*.