Using language models to probe the nature of “mild” island effects

Maho Takahashi (University of California, San Diego; mtakahas@ucsd.edu)

1. Background: Recent advancement in language models (LMs) have led researchers to probe LM’s syntactic knowledge [1-8] and apply the knowledge to discuss how humans acquire syntactic rules [9]. One of the most studied syntactic phenomena in this line of research is islands [10], a group of structures that disallow extractions out of them. The focus of our study is a relative clause (RC) island in Japanese, for two reasons: First, most research on the grammatical knowledge of LMs is about English, and we still know much less about whether LMs are capable of learning island constraints in any other languages. One study [11] examined the knowledge of Japanese RC island and found no evidence for it, but we will adopt a different paradigm that may be more sensitive to LM’s knowledge of islands. Second, while the violation of RC island in Japanese significantly decreases sentence acceptability [12], the effect of RC island seems rather small (or mild) compared with other cases of island violation. It has thus been proposed that there is no such thing as a mild island effect, and what looks like one is due to aggregating data, masking inter-item [13] or inter-participant [14] variability of acceptability judgments. Importantly, testing LMs have several advantages over testing human participants to investigate this issue, including their property that they do not undergo the satiation effect [15] unless they are fine-tuned (i.e., no inter-item variability). We therefore chose to test LMs on their knowledge of Japanese RC island and explore the nature of mild island effects.

2. Experiment: We tested GPT-2 models trained with Japanese texts of three parameter sizes [16]: Xsmall, Small, and Medium (37M, 110M, 336M). In accord with the previous studies assessing the grammatical knowledge of LMs, we calculated surprisal [17,18] for each word $S(w_k)$ upon seeing the word w_k given h_{k-1}, the hidden state after processing all the previous words in a sentence: $-\log_2 \mathbb{P}(w_k|h_{k-1})$. The test stimuli had 2x2 design, manipulating whether or not a sentence has a RC licensor and/or a gap (i.e., RC head noun and a gap in its original position

(1) *no RC-licensor, no gap (a: non-island, b: RC)*

a.

[[Gakusha-ga	suiri syousetsu-o	kai-ta]	koto-ga	saikin
professor-NOM	mystery novel- ACC	write-PST	fact-NOM	recently
syoten-de	tokusyuu-sa-re-ta]	koto-wa	hokorashii.	
book.store-at	feature-do-PASS-PST	fact-TOP	proud	

‘(I’m) proud of the fact [that the fact [that the professor wrote the mystery novel] recently got featured at a bookstore].’

b.

[[RC Gakusha-ga	__$_j$	kai-ta]	suiri syousetsu$_j$-ga	saikin	syoten-de
professor-NOM		write-PST	mystery novel- NOM	recently	book.store-at
tokusyuu-sa-re-ta]		koto-wa	hokorashii.		
feature-do-PASS-PST		fact-TOP	proud		

‘(I’m) proud of the fact [that the mystery novel$_j$ [RC that the professor wrote__$_j$] recently got featured at a bookstore].’

(2) *[+RC-licensor] [+gap] (a: non-island, b: RC)*

a.

[RC [__$_i$	suiri shoosetsu-o	kai-ta-koto]-ga	saikin
	mystery novel-ACC	write-PST-fact-NOM	recently
shoten-de	tokushuu-sa-re-ta]	**gakusha**$_i$-wa	hokorashige-da.
bookstore-at	feature-do-PASS-PST	professor-TOP	looks.proud-COP

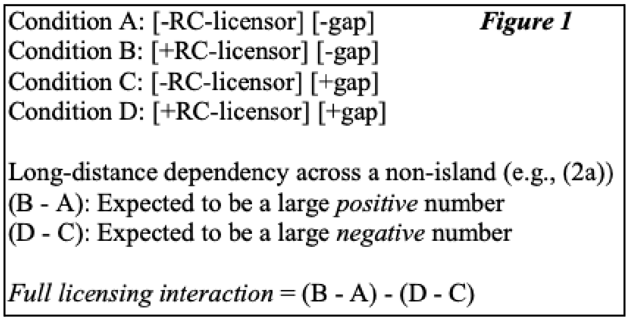
‘The professor$_i$ [RC who the fact that [__$_i$ wrote a mystery novel] was recently featured in a bookstore] looks proud.’

b.

[RC2 [RC1 __$_i$ __$_j$	kai-ta]	suiri shoosetsu$_j$-ga	saikin
	write-PST	mystery novel-NOM	recently
shoten-de	tokushuu-sa-re-ta]	**gakusha**$_i$-wa	hokorashige-da.
bookstore-at	feature-do-PASS-PST	professor-TOP	looks.proud-COP

‘The professor$_i$ [RC2 who the mystery novel$_j$ [RC1 that __$_i$ wrote __$_j$] was recently featured in a bookstore] looks proud.’

inside the embedded clause). Stimuli exemplifying two out of the four conditions ([-RC-licensor] [-gap], [+RC-licensor] [+gap]) are shown in (1) and (2). (1) exemplifies sentences that involve an embedded clause (a non-island complex noun phrase headed by *koto* 'the fact (that)' in (1a), a RC in (1b)) but no further gap or a RC licensor. In contrast, sentences in (2) involve further relativization of *gakusha* 'professor' out of the *koto*-clause (=2a) and the RC (=2b). The other two conditions ([+RC-licensor] [-gap], [-RC-licensor] [+gap]) are minimally different from (2); [+RC-licensor] [-gap] condition was derived from (2) by filling the sentence-initial gap __$_i$ with the noun identical to the head noun (*gakusha* in (2)). As Japanese permits the so-called gapless RCs (RCs that do not have a gap corresponding to the head noun), filling the gap with any nouns other than the one identical to head noun could make the RC parsed as gapless, instead of as the RC with a filled gap. [-RC-licensor] [+gap] condition was derived by deleting ***gakusha***$_i$. LMs' knowledge about long-distance dependency and the RC island were probed by comparing the mean surprisal of the critical region (grayed) in the four conditions, using the metric called *licensing interaction* [8,9]. As the formula in Figure 1 illustrates, a positive licensing interaction value is indicative of the model's knowledge of long-distance dependency. Crucially, if LMs are aware of island constraints, this value is predicted to decrease for long-distance dependency across an island. This is because LMs would be less likely to expect a relationship between the RC-licensor and the gap if the latter is inside an island.



Condition A: [-RC-licensor] [-gap] ***Figure 1***
Condition B: [+RC-licensor] [-gap]
Condition C: [-RC-licensor] [+gap]
Condition D: [+RC-licensor] [+gap]

Long-distance dependency across a non-island (e.g., (2a))
(B - A): Expected to be a large *positive* number
(D - C): Expected to be a large *negative* number

Full licensing interaction = (B - A) - (D - C)

Results: As the top row of Table 1 shows, our results are in line with the findings of previous studies that LMs can learn long-distance dependency between a licensor and its gap, as positive licensing interaction values indicate. The bottom row shows that the licensing interaction value did not decrease for dependency across the RC island in Xsmall model, suggesting that the model has not learned the island constraint. In contrast, there is a small yet noticeable decrease in licensing interaction values among Small and Medium models (-0.60 and -0.29).

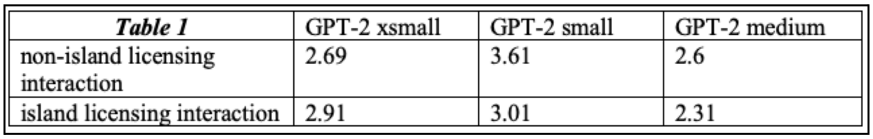


Table 1	GPT-2 xsmall	GPT-2 small	GPT-2 medium
non-island licensing interaction	2.69	3.61	2.6
island licensing interaction	2.91	3.01	2.31

3. Discussion & Conclusion: The decrease in the licensing interaction for some of the Japanese GPT-2 models we tested (Small and Medium) indicates that they are aware of the constraints on long-distance dependency, namely that dependency cannot be formed across a RC island. Furthermore, only the slight decrease emulates the pattern of acceptability judgments made by humans; as noted in Section 1, the violation of RC island in Japanese seems to have a significant yet mild effect on acceptability. Given the properties of LMs that their pattern of judgments is less prone to variability, it is now likely that the mild RC island effect exhibited by human participants is genuine, rather than due to aggregating judgments across participants and items.

References: **[1]** Linzen et al. (2016); **[2]** Lau et al. (2017) **[3]** Bernardy & Lappin (2017); **[4]** Kuncoro et al. (2018); **[5]** Gulordava et al. (2018); **[6]** Futrell et al. (2018); **[7]** Marvin and Linzen (2018); **[8]** Wilcox et al. (2018); **[9]** Wilcox et al. (2021); **[10]** Ross (1967); **[11]** Takahashi (2023); **[12]** Takahashi & Goodall (2021); **[13]** Kush et al. (2019); **[14]** Fukuda et al. (2023); **[15]** Snyder (2000); **[16]** rinna Co., Ltd. (2021); **[17]** Hale (2001); **[18]** Levy (2008)