The success of Neural Language Models on syntactic island effects is not universal: strong *wh*-island sensitivity in English but not in Dutch

Michelle Suijkerbuijk, Naomi Tachikawa Shapiro, Peter de Swart, Stefan Frank
Centre for Language Studies, Radboud University

Background: A much-debated question in linguistics is how humans acquire grammatical knowledge: are we born with a language-specific learning capacity, or can we learn language from input alone? The recent introduction of neural language models (NLMs) is greatly influencing this debate. NLMs learn solely from their input in combination with their inductive biases, and thus without any built-in linguistic representations. If these networks can learn specific grammatical phenomena similarly to humans, this suggests that these phenomena can, in principle, be learned based on input alone. Recent research has looked at the learnability of one of the most studied phenomena in experimental syntax, syntactic island effects (see example in Table 1), to investigate whether NLMs are sensitive to island violations (e.g., [1]). Syntactic islands are an ideal test bed because they rarely occur in training data and NLMs do not have built-in linguistic knowledge to fall back on. Research has mostly shown successful results: NLMs seem able to model island effects in English. Yet, the behaviors or NLMs are almost never compared to human data and are almost exclusively researched in English. This makes it difficult to claim that NLMs can model island effects in ways that are comparable to humans. The present study addresses these gaps by incorporating data from human experiments and by looking beyond English.

Method: We make two important improvements on earlier work. First, we present NLMs and human participants with the same sentences (Table 1).[1] By collecting both model-assigned sentence probabilities and participant acceptability judgments, we directly compare whether the model represents island sensitivity similarly to humans. Second, we take this approach beyond English and compare NLM and human behavior in both English and Dutch, since the languages, though related, differ in their word order (SVO vs. SOV).

Results: The results for English and Dutch are shown in Figure 1. Figure 1 (top two plots) shows that the strong *wh*-island sensitivity of NLMs in English is replicated and that this sensitivity is comparable to that of English participants: the NLMs and the participants show the same patterns in their results. The same cannot be said for Dutch, however (bottom two plots of Figure 1). While the Dutch participants showed a strong sensitivity to *wh*-island violations comparable to English participants and models, the Dutch NLM did not.

Conclusion: NLMs are not successful in all languages (yet) (e.g., Dutch), so more cross-linguistic research is necessary before NLMs can be claimed to bear on the human capacity for grammar learning.

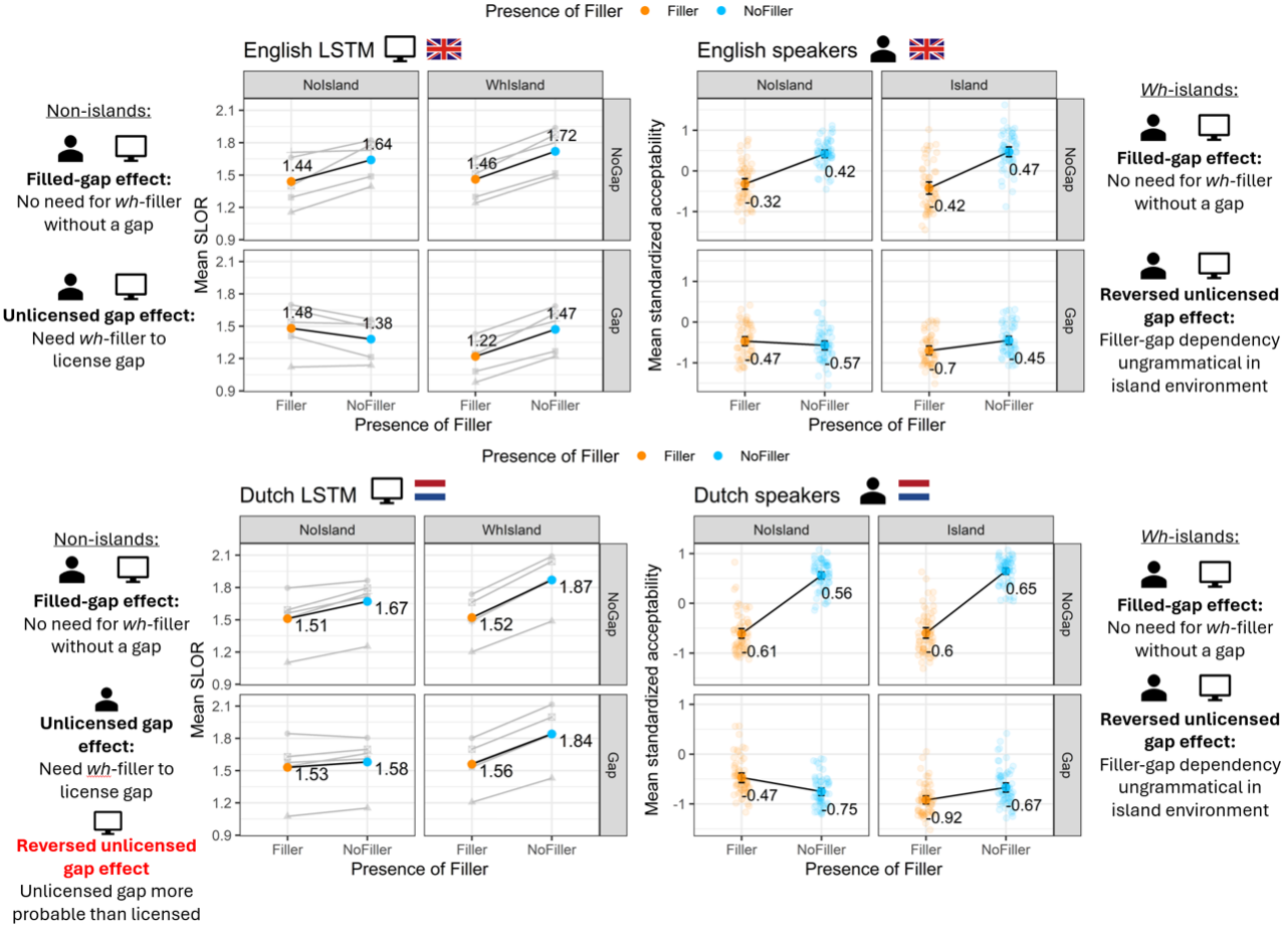
[1] We trained five Long Short-Term Memory networks with one recurrent layer per language. All model were trained on ~9M sentences from either the NLCOW corpus in Dutch or the ENCOW corpus in English, both consisting of individual sentences collected from the World Wide Web.

Table 1. *Example sentences used in the Dutch (NL) and English (EN) experiments, crossing the factors* PRESENCE OF GAP *('cookies' vs. _ (a gap); indicated in orange),* PRESENCE OF FILLER *('that' vs. 'what'; indicated in blue) in non-islands and* wh*-islands (indicated in red between curly brackets).*

Gap?	Filler?		Example sentence in non-island and ***wh*-island** configuration
No	No	NL	*Ik weet dat jij {denkt dat/betwijfelt of} de bakker koekjes maakt in de bakkerij.*
			I know that you {think that/doubt if} the baker cookies makes in the bakery
		EN	I know that you {think that/doubt whether} the baker makes cookies in the bakery.
No	Yes	NL	**Ik weet wat jij {denkt dat/betwijfelt of} de bakker koekjes maakt in de bakkerij.*
			I know what you {think that/doubt if} the baker cookies makes in the bakery
		EN	*I know what you {think that/doubt whether} the baker makes cookies in the bakery.
Yes	No	NL	**Ik weet dat jij {denkt dat/betwijfelt of} de bakker _ maakt in de bakkerij.*
			I know that you {think that/doubt if} the baker GAP makes in the bakery
		EN	*I know that you {think that/doubt whether} the baker makes _ in the bakery.
Yes	Yes	NL	*Ik weet wat jij {denkt dat/*betwijfelt of} de bakker _ maakt in de bakkerij.*
			I know what you {think that/doubt if} the baker GAP makes in the bakery
		EN	I know what you {think that/*doubt whether} the baker makes _ in the bakery.

Note. The Dutch and English sentences only differ in the object-verb order in the embedded sentence (*koekjes maakt* vs. 'makes cookies').

Figure 1. *Mean standardized acceptability judgements (right plot) and mean Syntactic Log-Odds Ratio value (i.e., average, frequency- and length-corrected surprisal; left plot) for every combination of* PRESENCE OF GAP *and* PRESENCE OF FILLER *within non-islands (top and bottom left) and* wh*-islands (top and bottom right) for English (top plots) and Dutch (bottom plots). Error bars represent 95% confidence intervals.*



References

[1] Wilcox, E. G., Futrell, R., & Levy, R. P. (2024). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, *55*(4), 805-848.