Gone without a Trace: Why 5 LMs Fail to Track Non-Local Dependency Resolution

Previous work has shown that probability estimates generated by language models (LMs) are unable to account for human data in the processing of complex sentences, including garden-paths [1] and non-local filler-gap (FG) dependencies [2]. One explanation is that structures such as garden-paths involve complex revision processes which go beyond next-word prediction [2]. However, even when a partial analogue to human revision is provided by LMs with limited parallelism, the magnitude of human difficulty remains unaccounted for [3]. This raises the possibility that humans may form more detailed syntactic representations, whereas LMs rely primarily on surface-level heuristics to generate predictions. Attempts to train LMs to explicitly represent constituent structure only partially improve performance on multiclausal FG dependencies [4]. We hypothesise that this is because human FG resolution relies on traces to track the dependency, particularly non-locally, and explore whether training LMs on corpora annotated for traces better predicts human data.

Our human data comes from a novel implementation of the Maze Task, which demonstrates that parsers have a stronger preference for subject compared to object gaps in non-local vs. local dependencies [2]. This pattern is confirmed by corpus frequencies from both the Penn Tree Bank (PTB) as well the scrape of English Wikipedia from [5], parsed for constituent structure using the Berkley Neural Parser (**Table 1**). Using the code and corpora at https://osf.io/r2t6d/overview?view_only=f4953c834f1e407e9d5541e8b5854671, we trained LSTM models from scratch using best hyper parameters from [5] and fine-tuned more powerful GPT-2 Small models [6]. Two different corpora were used: 1. the original English Wikipedia scrape, and 2. the same corpus annotated for traces (*T*), which are inserted as lexical items into the strings. The LMs were trained to predict the next word (including *T*); additionally, we trained an LSTM using the annotated corpus to binary predict whether the next token was *T* or not. This was to have a simplified training objective, as well as to ensure that including *T* in the corpus did not affect LM word order preferences. The 5 trained models are summarised in **Table 2**. We used the LMs to generate probability estimates for either the word cueing the subject gap continuation in [2] relative to the object continuation, or *T* relative to no *T*, depending on the training corpus (unannotated/**annotated**), e.g. '*The girl who the teacher remarked should / ***T***.* '.

Probability estimates from all models failed to predict the human data across the two environments. However, models that were trained on the corpus without explicit traces predicted the subject-gap preference in the local environment, but not non-locally (**Table 2**, see OSF for full statistical results). We suggest that, although structural predictions are critical for FG dependency resolution (e.g. Active Filler Strategy [7]), humans likely predict a word which is compatible with a trace (i.e. the verb from [2]), rather than directly predicting a gap. LMs may not predict gap preferences non-locally, due to the extremely low corpus frequencies of these structures (**Table 1**). If the limited evidence for gaps in non-local structures causes the LM to 'lose track' of the dependency, we should expect there to be lower Surprisal for an object 'filled-gap' effect non-locally compared to locally, which was confirmed (**Figure 1**). Non-local dependencies are also infrequent in the human input, however, displaced arguments (i.e. fillers) have been argued to receive a privileged status in human memory until thematically interpreted [7]. This privileged status in memory allows the human parser to more easily generalise FG dependencies over structural distances. Future work should look at whether explicitly privileging and tracking the filler in LMs improves prediction of human data.

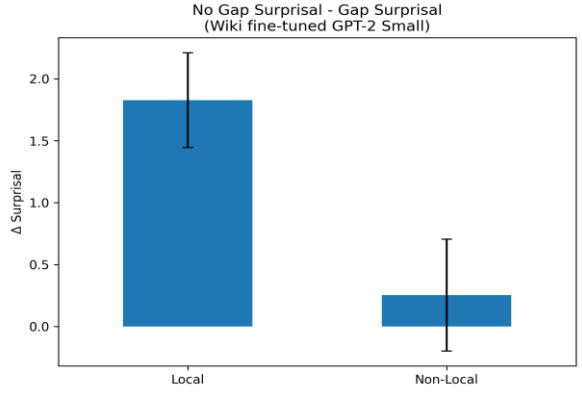
Table 1: Subject Gap Frequency in Human Data Locally/Non-Locally when Controlling for Sentence-Initial Position

	Data	Example Sentence from [4] showing Subject/Object Continuation / Tregex Query showing Prefix and **Subject Gap Continuation**	Subject Gaps
Local	Maze	*The teacher remarked that the girl should be rewarded … / the department should reward …*	66% s.e.=1.9
	PTB	*(@NP<!-NONE-.(@VP<(@SBAR<(@S<(@NP<(@SBAR<(@WHNP $(@S<1**(@NP</-NONE-<*T*/$(@VP[!<TO])**))))))))$/\./)>,@S*	77% **83**/108
	Wiki	S[!$-- __]<1(NP$..(VP<(SBAR<(S<(NP<(NP$(SBAR [!<WHPP&!<WHADVP&!<IN]&<(S**[!<NP]&<(VP[!<TO])**))))))))	70% **1755**/2510
Non-local	Maze	*The girl who the teacher remarked should be rewarded … / the department should reward …*	90% s.e.=1.1
	PTB	*(@NP<!-NONE-<(@SBAR<(@WHNP$(@S<((@NP<!-NONE-) $(@VP<(SBAR<(@S<1**(@NP</-NONE-<*T*/$(@VP[!<TO])**))))))) $/\./)>, @S*	100% **7**/7
	Wiki	S[!$-- __]<1(NP<(NP$(SBAR[!<WHPP&!<WHADVP&!<IN]&<(S< (NP$..(VP<(SBAR[!<IN&!<WHADVP]&<(S**[!<NP]&<(VP[!<TO])**))))))))	84% **41**/49

Table 2: Model architectures and training, including predictability of human data using: glmer (Continuation~frequency+Animacy*Locality*Gap… and Locality/Gap …) with full RE for Participants and Items. Animacy effect orthogonal to current research

LM	*T*	Objective	Valid PPL	Locality *Gap	Non-local/ Local
LSTM	x	Language Modeling	51.0	***p=.042***	*NL: p=.932* ***L: p=.008***
	✓	Language Modeling	50.2	*p=.351*	N/A
	✓	Binary Prediction	N/A	*P=.273*	N/A
GPT-2 (S)	x	Language Modeling	23.5	***p=.002***	*NL: p=.263* ***L: p=.001***
	✓	Language Modeling	23.2	*p=.080*	N/A

Figure 1: Bar chart showing reduced object 'filled-gap' effect non-locally using Surprisal from our LM that best predicts human data. Error bars represent S.E.M.



References

[1] Huang, K. … (2024) Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty, *JML*, 137

[2] Nentcheva, L. … (2025). *Surprise!* LM-Generated Surprisal Values Cannot Fully Account for Human Parsers' Preferences in Relative Clause Comprehension. Talk at: *38th HSP*

[3] Timkey, W. & Linzen, T. (2025). Language model surprisal underpredicts garden path effects even with limited syntactic parallelism, Poster at: *38th HSP*

[4] Wilcox, E. … (2019). Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. *Proc. 2019 Conf. NA Chpt. of Assoc. for Comp. Ling*

[5] Gulordava, K. … (2018). Colorless Green Recurrent Networks Dream Hierarchically. In *Proc. 2018 Conference of the NA Chapt. of the Assoc. for Comp. Ling.*

[6] Radford, A. … (2019). Language Models are Unsupervised Multitask Learners. OpenAI

[7] Frazier, L. & Clifton, C. (1989). Successive cyclicity in the grammar and parser. *Lang. Cog.* 4(2)