Processing Gaps: Explaining Human-Model Misalignment in Filler-Gap Dependency Parsing

Liliana Nentcheva[1], Sebastian Schuster[2], Donald Dunagan[3] and Andrea Santi[1]

[1]University College London (UCL); [2]University of Vienna; [3]Johns Hopkins University

Corresponding author: lily.nentcheva.15@ucl.ac.uk

Recent theories of human parsing emphasise the contribution of both contextual (un)predictability and memory demands to a word's processing difficulty [1,2]. Temporarily ambiguous sentences which terminate in the less frequent continuation generate reading time (RT) slowdowns of much greater magnitude than what is predicted by language-models' (LM) estimates of next-word probability [1]. One explanation is that limited memory encourages human parsers to commit to a single structural interpretation, resulting in costly reanalysis when disconfirmed by later input [1]. Although LMs can represent multiple readings of ambiguous sentences [3], even LMs with limited parallelism fail to account for human difficulty in these 'garden-path' structures [4]. This raises the possibility that corpus data is not accurately reflected in the model's learned probabilities of competing continuations [4], perhaps due to an impoverished syntactic representation. We ask if human parsing preferences are predicted by LM-probability estimates in structures with high memory demands that do not require revision.

Human data is provided from a novel implementation of the Maze Task [5], where participants were required to choose between a relative clause (RC) continuation which unambiguously signals a subject (e.g. *should*) or object (*the*) gap. The selection was made either at a non-local choice-point, within a complement clause of the RC verb (Exp.1, n=76) or locally, following the relativiser (Exp.2, n=62) (**Table 1**). Humans demonstrated a strong subject gap preference non-locally (0.91) and, to a lesser extent, also in the local condition (0.66) [5]. This aligns with frequency counts in the Penn Tree Bank (PTB): using a Tregex search for constituent structure (without specifying lexical items), we found a heightened subject gap probability non-locally (1.00) compared to locally (0.83) (**Table 2**). In the local condition, all other trees included object gaps, such that other available parses are too infrequent to constitute serious competitors. Surprisal [6] estimates were subsequently generated for the words signaling the subject vs. object gap continuations (*should* and *the*) using GPT-2 (small). To facilitate linking without transformation assumptions (as previously relied on, e.g. [1]), we converted the Surprisal values into the relative probability of the subject gap parse (pSgap) (**Equation 1**), to be on the same scale as our human data. pSgap predicted the subject-gap preference locally (p<.001), but did not predict the strength of the preference non-locally (p=.426), demonstrating a similar probability of the subject gap continuation across dependency lengths (**Table 3**).

It is expected that the human preference and corpus data align in showing a higher subject gap probability non-locally, given that human parsers are motivated to integrate the filler early across both comprehension and production to reduce memory demands [7]. Failure of the LM to predict the non-local data could result from human-like memory constraints not being faithfully reflected in the model architecture, or the absence of an explicit syntactic representation of the input. While prior attempts to combine syntactic and lexical surprisal have shown limited success in approximating human data for complex sentences [8], including movement dependencies with multiple embeddings [9], these models were trained on relatively 'superficial' representations of syntax. Movement dependencies rely on hierarchical relations and may require an LM which explicitly represents gaps. A promising direction for future work is to model these effects using an LM provided with more sophisticated syntactic supervision to disentangle the contribution of structural representations from memory constraints.

Table 1. *Example items; **bolded** word indicates where Surprisal and parser preferences were measured in RCs <u>with</u> (Exp. 1) and without (Exp. 2) an intervening clause*

Experiment	*Pre-Decision Point*	***Subject/Object** Gap Continuation*	*Matrix Clause*
1. Non-local	*The girl* who <u>the teacher remarked</u>*	***should** be rewarded by the department*	
		***the** department should reward*	*had surpassed all expectations*
2. Local	*The teacher remarked that the girl who*	***should** be rewarded by the department*	
		***the** department should reward*	

Equation 1. *Probability of a subject gap parse (pSgap) calculated from Surprisal values*

(i) $\text{Surprisal}_{subj/obj} = -\log P(w_{aux/det}|1\ldots w_{aux/det-1})$

(ii) $P(w_{aux/det}|w1\ldots w_{aux/det-1}) = 2^{-\text{Surprisal}_{subj/obj}}$

(iii) $\text{pSgap} = 2^{-\text{Surprisal}_{subj}}/(2^{-\text{Surprisal}_{subj}}+2^{-\text{Surprisal}_{obj}})$

Table 2. *Corpus analysis on the PTB corpus of ~50,000 trees using the following Tregex queries*

	Query Term	*Matches*
Non-local(Exp.1)	`(@NP<!-NONE- <(@SBAR<(@WHNP$(@S<(( @NP<!-NONE-) $(@VP<(SBAR<(@S<1(@NP< /-NONE-/)))))))) $/\./)>, @S`	*Prefix: 7* ***Subject Gap: 7*** <u>Probability: 1.00</u>
Local (Exp.2)	`(@NP<!- NONE-.(@VP<(@SBAR<(@S <(@NP<(@SBAR<(@WHNP$( @S<1(@NP</-NONE-)))))))) $/\./)>, @S`	*Prefix: 108* ***Subject Gap: 90*** <u>Probability: 0.83</u>

Table 3. *Results based on simple effects model: glmer(Continuation~Animacy+Experiment/pSgap+(1+Animacy|Participant)+(1+Animacy|Item)*

*Animacy manipulation tangential to current research

Fixed Effects	β	Std. Error	*z*-value	*p*-value
Animacy*	0.281	0.074	3.782	< .001
Experiment	1.722	0.220	7.837	< .001
Exp.1: pSgap	-0.386	0.485	-0.795	.426
Exp.2: pSgap	1.815	0.328	5.540	< .001

<u>References</u>

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

[2] Futrell, R… (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cogn Sci*. 44(3)

[3] Hanna, M. & Mueller, A. (2025). Incremental Sentence Processing Mechanisms in Autoregressive Transformer Language Models. *Proc. 2025 Conference of Chpt. of the Assoc. for Comp. Ling.* 1

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

[5] Nentcheva, L. & Santi, A. (2024). Parsers Predict Subject Gaps Even for Inanimate Fillers: Evidence from an Innovation to the Maze Task. *Poster at AMLaP, 2024*

[6] Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *2nd Meeting NA Chapt. of the Assoc. for Comp. Ling.*

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

[8] Arehalli, S… (2022). Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. *Proc.26th Conference on Comp. Nat. Lang. Learning*

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