Bounded Rationality in Hindi Word Order: Evidence from Human Speech and LLMs

Manvendra Singh[1] (manvendrasinghemail@gmail.com), Sidharth Ranjan[1]
[1]School of Artificial Intelligence and Data Science, Indian Institute of Technology Jodhpur

Dependency Length Minimization (DLM) posits that speakers keep syntactically related words close together to minimize cognitive load within sentences [1,2]. However, achieving such globally optimal dependency arrangements is computationally expensive as it involves search over a large combinatorial space and consumes substantial working memory. Recent research demonstrated that in written Hindi, a relatively free word-order, Subject-Object-Verb (SOV) language, humans bypass global optimization for word order preferences [3]. Instead, they employ a computationally cheaper "least-effort" bounded-rationality heuristic: placing a short preverbal constituent immediately adjacent to the main verb. This configuration concurrently shortens the length of all preverbal dependencies and minimizes the total dependency length of the sentence without the need to explore the search space of constituent orders [4]. The current work extends this framework to two novel domains: spontaneous human speech, where real-time cognitive constraints are strict, and machine-generated texts using Large Language Models (LLMs), which lack human biological memory limits. We investigate whether the "least-effort" heuristic persists under extreme cognitive load and if it is replicated by large language models.

Methodology: For human evaluation, we extracted spontaneous speech transcripts from the spoken section of the Hindi EMILLE corpus [5]. The spoken dataset comprised three diverse sub-genres: Afternoon (casual conversations), Asiannet (news broadcasts), and Monologue. For machine evaluation, we constructed a controlled dataset by prompting six state-of-the-art LLMs: Sarvam-30B, Qwen-3-32B, ChatGPT-5.5, Llama-3.1-8B, Gemma-3-4B, and Gemma-3-12B. Crucially, we employed a cross-lingual prompting strategy (English instructions for Hindi outputs) to prevent syntactic mimicry and ensure independent generation. After this, all sentences were parsed using the Stanza dependency parser [6], following the universal dependency annotation scheme. We selected dependency trees with two or more preverbal constituents and generated all grammatically valid counterfactual variants via random permutation of preverbal constituents. Next, we conducted corpus analysis, where we examined the distributions of length of preverbal constituents in the corpus sentences. We then deployed logistic regression models to classify the word order attested in the corpus against the artificially generated variants using two predictors, *viz.,* global dependency length (Total DL) and length of the preverbal constituent immediately preceding the main verb (CL Last), denoting the least effort strategy.

Results and Discussion: Our logistic regression models yielded significant negative coefficients across all datasets, with the local Last Constituent Length exhibiting a substantially larger negative effect size than Total DL (see Table 1), confirming the *least effort strategy*. Additionally, despite the fragmented syntax of speech, the local least-effort strategy consistently outperforms global DLM optimization in terms of prediction accuracy, not just in human speech but also across other LLM-generated datasets. For instance, in the Afternoon spoken sub-genre, the local heuristic (CL Last) predicted word order with 63.07% accuracy versus 59.16% for global DLM (Total DL). Strikingly, across all evaluated variants, modern LLMs successfully replicate this bounded rationality pattern. Across all models, CL Last significantly outperformed Total DL (e.g., 77.71% vs. 63.90% in Llama-3.1-8B). Furthermore, combining the lengths of the last two constituents pushed predictive accuracy even higher (79.13% for Llama-3.1-8B; 63.44% for Afternoon) over and above the global DLM predictor, confirming that both biological and artificial systems selectively optimize the immediate preverbal syntactic window.

Conclusion: Our results suggest that the "least-effort" heuristic is robust for predicting constituent-ordering choices in spontaneous Hindi speech. Additionally, we conjecture that despite their capacity for global optimization, LLMs seem to operationalize Hindi syntax using a similar heuristic. Future work will investigate whether LLMs implicitly learn human-like cognitive limitations from their training data. We also plan to extend this framework to diverse language families, investigate whether the heuristic also governs real-time language production, and develop a mechanistic account of cognitive processes underlying DLM.

Table 1: Classification Accuracy and Regression Coefficients: Comparing predictive accuracy and regression estimates of Global Dependency Length (Total DL) vs. the local "Least-Effort" (CL Last) heuristic across human speech and LLMs.

Dataset / Model (#Ref, #Var data points)	(a) Classification accuracy (%)			(b) Regression coefficients		
	Total DL	CL Last	Total DL + CL Last	Total DL	CL Last	Total DL + CL Last
EMILLE Spoken: Afternoon (2742, 58983)	59.16	63.07	63.44	-0.56	-0.72	-0.19, -0.60
EMILLE Spoken: Asiannet (2352, 46674)	60.49	62.06	62.56	-0.87	-0.99	-0.48, -0.75
EMILLE Spoken: Monologue (791, 11019)	42.84	46.58	42.14	-0.25	-0.12	-0.28, 0.05
LLM: ChatGPT-5.5 (4848, 37721)	64.39	70.68	71.71	-0.75	-1.31	0.03, -1.34
LLM: Gemma-3-4B (5495, 138767)	65.19	76.11	77.15	-0.75	-1.95	0.16, -2.06
LLM: Gemma-3-12B (6242, 129221)	64.16	76.21	78.17	-0.72	-2.1	0.20, -2.24
LLM: Llama-3.1-8B (7115, 225695)	63.9	77.71	79.13	-0.67	-2.04	0.33, -2.27
LLM: Qwen-3-32B (1663, 28400)	59.85	70.8	72.31	-0.42	-1.38	0.31, -1.59
LLM: Sarvam-30B (2491, 48983)	58.79	73.77	78.19	-0.37	-1.68	0.62, -2.15

Columns (a) report the predictive classification accuracy, while columns (b) present the underlying regression coefficients for each model (all reported coefficients are highly significant with $p < 0.001$). ***Table Glossary:*** *Total DL: Total Dependency Length (global sentence optimization); CL Last: Constituent Length of the immediate preverbal phrase (local "least-effort" optimization); LLM: Large Language Model; #Ref, #Var: Reference sentences, computationally generated grammatical Variants.*

Table Analysis: The Monologue dataset reveals that extreme working memory burden causes local least-effort planning to fragment (hybrid accuracy drops to 42.14%). Conversely, LLMs exhibit a striking coefficient reversal in the hybrid model: while strongly penalizing local lengths (e.g., -2.27 in Llama-3.1-8B), their global Total DL coefficients become positive (+0.33). This proves artificial networks strictly optimize the immediate preverbal slot, but do not inherently penalize longer global dependencies.

References

[1]. E. Gibson, "Dependency locality theory: A distance-based theory of linguistic complexity," Image, language, brain: Papers from the first mind articulation project symposium, Cambridge, MA: MIT Press, 2000.

[2]. Futrell, R., Mahowald, K., & Gibson, E. Large-scale evidence of dependency length minimization in 37 languages. Proceedings of the National Academy of Sciences, 112(33), 10336-10341, 2015

[3]. Ranjan, S., & von der Malsburg, T. A Bounded Rationality Account of Dependency Length Minimization in Hindi. In Proceedings of the 45th Annual Meeting of the Cognitive Science Society (CogSci), 2023.

[4]. H. A. Simon, "Models of bounded rationality," Economic Analysis and Public Policy, MIT Press, Cambridge, Mass, 1982.

[5] Anthony McEnery, Paul Baker, Robert Gaizauskas, and Hamish Cunningham. Emille: Building a corpus of south asian languages. In Proceedings of the International Conference on Machine Translation and Multilingual Applications in the New Millennium. 2000.

[6]. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages, 2020.

LLM Prompting Strategy and Sentence Generation: To ensure the LLMs generated pure Hindi without any English code-mixing, we constrained them using a strict zero-shot prompt. To prevent topical bias and mirror the domain diversity of the EMILLE corpus, we dynamically inserted {topic} and {sub_genre} variables into these prompts. These variables were drawn from a curated list of over 50 diverse scenarios (e.g., News, Science, Government). The exact prompts used for generation were:

[SYSTEM_PROMPT = "You are a professional Hindi writer and language expert in the field of {sub_genre}."
USER_PROMPT = """Task: Write a natural paragraph of exactly {N} sentences about the given topic.
Topic: {topic}
Sub-genre: {sub_genre}
Rules:
- Write exactly {N} sentences, no more, no less.
- The text must be completely natural and flowing, exactly as humans typically write articles or news reports.
- Use ONLY natural Hindi in the Devanagari script. Do NOT use any English words, Roman letters, or brackets (e.g., no English words in parentheses).
- Write all sentences together in a single paragraph block.
- VERY IMPORTANT: Do NOT provide any title/heading, bold text (**), lists, numbers, or additional commentary. Start the paragraph directly without any title."""]

Typological Features and Syntactic Flexibility of Hindi: Hindi is an Indo-Aryan language characterized by a default Subject-Object-Verb (SOV) word order and a rich system of postpositional case-marking. Because these case markers bind directly to nominals, the preverbal constituents can be freely permuted without altering the core propositional meaning of the sentence. This extreme syntactic flexibility makes Hindi an ideal testbed for evaluating dependency length minimization (DLM), as it allows us to generate a massive evaluation space of valid counterfactual variants via random permutation.

Example of Preverbal Permutation:

Reference : [kaidiyon=ne] [kaanistebalon=se chheene hathiyaar] **[vahiin]** gira diye.
Gloss: [Prisoners=ERG] [constables=ABL snatched weapons] [there] dropped.PFV AUX.
Translation: *The prisoners dropped the weapons snatched from the constables right there.*
Variant-1: [kaidiyon=ne] [vahiin] [kaanistebalon=se chheene hathiyaar] gira diye.
Variant-2: [kaanistebalon=se chheene hathiyaar] [kaidiyon=ne] [vahiin] gira diye.
Variant-3: [kaanistebalon=se chheene hathiyaar] [vahiin] [kaidiyon=ne] gira diye.
Variant-4: [vahiin] [kaidiyon=ne] [kaanistebalon=se chheene hathiyaar] gira diye.
Variant-5: [vahiin] [kaanistebalon=se chheene hathiyaar] [kaidiyon=ne] gira diye.

(Note: In the reference sentence, the human speaker naturally satisfies the "least-effort" heuristic by placing the absolute shortest constituent, 'vahiin' (1 word), in the immediate preverbal slot, rather than holding the heavier 2-word (Variant-1) or 4-word constituent (Variant-2) right next to the verb. By computationally shuffling these bracketed preverbal constituents, we generated all variants used in our logistic regression to track this cognitive preference).