# Limits of Dependency Length Minimization

Sidharth Ranjan and Titus von der Malsburg
University of Stuttgart

Dependency locality in the form of dependency length minimization (DLM) has been demonstrated as an explanatory principle behind word order preferences in natural languages (Futrell et al., 2020). This principle seeks to keep any pair of linked head-dependent words in a dependency tree as *close* as possible in their linear order within a sentence due to efficiency factors from limited memory capacity. It is unclear, meanwhile, how much DLM is employed in a specific language. Furthermore, the cognitive processes driving the minimization of dependency length is unknown. Following a recent study by Ranjan and von der Malsburg (2024), we hypothesize that placing a short preverbal constituent next to the main verb explains constituent ordering decisions better than the global minimization of dependency length across SOV languages. We refer to it as "least-effort" strategy, as it reduces the dependency lengths between the verb and all its preverbal dependencies but does so in a cost-effective manner by streamlining the search space of possible constituent orders. We substantiate our hypothesis using large-scale corpus evidence from Universal Dependency Treebank (Zeman et al., 2022). Finally, we argue that our findings can be situated within the frameworks of *good-enough* account of language processing (Ferreira et al., 2002), and *bounded rationality* in decision-making (Gigerenzer et al., 2011), where *fast-but-frugal* heuristics hold precedence over extensive searches for optimal solutions.

**Method.** Our dataset includes sentences from the seven major SOV languages in the UD treebank: *Basque, Hindi, Japanese, Korean, Latin, Persian, and Turkish*. For each natural sentence (reference; 'ref') in the corpus, representative of human preferred choice, a large number of counterfactual variants ('var') were automatically created by randomly permuting the preverbal constituents in the sentence whose head was directly dependent on the root verb in the dependency tree (see Ex. 1 for an illustration with four preverbal constituents $C_i$ in a sentence). We then examined the distribution of the length of preverbal constituents and dependency lengths within a sentence. Thereafter, we tested our main hypothesis by deploying these two predictors in a logistic regression model to distinguish reference sentences from the generated variants.

**Results.** As the preverbal constituents approach the main verb, the global DLM would predict a gradual decline in their lengths. On the other hand, the least-effort strategy would expect optimization mostly on the preverbal constituent next to the main-verb. Fig. 1 validates the prediction across SOV languages, implying that sentences in the natural corpus show a preference for either optimizing the length of preverbal constituent next to the main verb or at least prefer it. Next, if speakers employ least-effort strategy, the length of constituent closest to the verb ('CL Last') should be better at predicting correct choices *i.e.,* corpus reference sentences ('ref') against generated variants ('var') than total dependency length ('Total DL'). Fig. 2 presents the summary of our dataset and Table 1 the results of our models. Aligned with our prediction, we found that 'CL Last' consistently outperformed 'Total DL' in predicting corpus reference sentences, in terms of classification accuracy (% of correctly predicted reference sentences, 'ref'), except Basque and Japanese. 'Total DL' and 'CL Last' gave same outcome for $\sim$70% cases across SOV languages with success cases (%Correct) indicated in parenthesis. The percentage of different outcomes can be inferred from the same column (100 minus %Same). Further, adding 'CL Last' feature over a baseline model with 'Total DL' feature, induced a significant increase in the accuracy ($p < 0.001$ using McNemar's test) for all SOV languages, including Basque and Japanese.

**Conclusion.** Overall, our findings show that SOV speakers minimize dependency length by considering only a limited search space of constituent orders, likely to conserve resources within the bounds of rationality and good-enough processing.

(1) a. $C_1$ $C_2$ $C_3$ $C_4$ $V$

b. $C_3$ $C_2$ $C_1$ $C_4$ $V$

c. $C_4$ $C_1$ $C_2$ $C_3$ $V$

d. $C_1$ $C_2$ $C_4$ $C_3$ $V$

**Constituent length** (CL) was calculated by counting words in a constituent ($C_i$), and the **total dependency length** (Total DL) of a sentence by summing the distances (in terms of words) between all head-dependent pairs in a dependency tree.

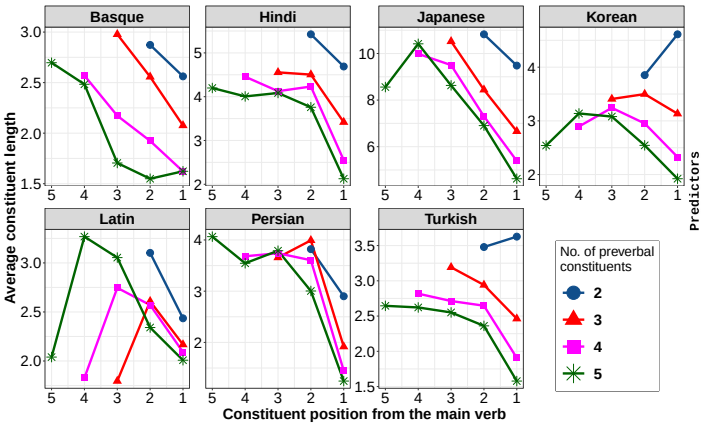


Figure 1: Average length of preverbal constituents in the corpus sentences containing only-2 to only-5 preverbal constituents

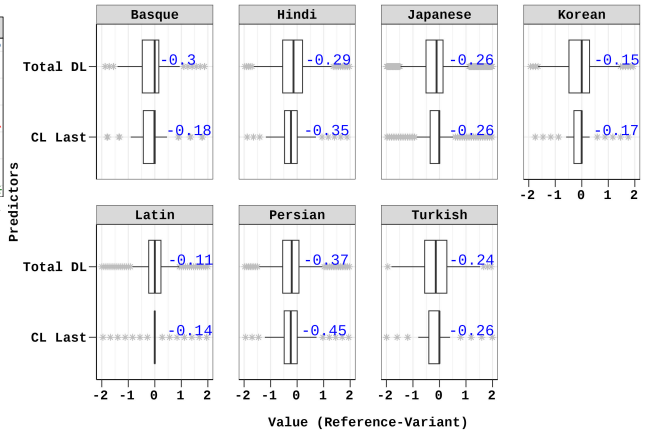


Figure 2: Summary of dataset denoting difference between predictor values of reference and the paired variant sentences; Mean difference values annotated inside the subplot

| Language | CL Last | Total DL | Total DL + CL Last | %Same (%Correct) |
|---|---|---|---|---|
| | Classification Accuracy (%) | | | Prediction (CL Last *vs.* Total DL) |
| Basque | 55.07 | 61.71 | 62.01 | 80.40 (48.59) |
| Hindi | 69.49 | 63.39 | 69.23 | 75.03 (53.97) |
| Japanese | 62.80 | 63.09 | 64.36 | 75.47 (50.68) |
| Korean | 56.92 | 55.11 | 56.44 | 76.11 (44.08) |
| Latin | 51.48 | 48.51 | 49.55 | 79.60 (39.79) |
| Persian | 74.57 | 69.04 | 75.17 | 68.69 (56.16) |
| Turkish | 61.72 | 60.00 | 62.02 | 77.44 (49.58) |

Table 1: Classification accuracy (%) of various models (10-fold cross-validation) with constituent length of last preverbal constituent (CL Last) and total dependency length (Total DL) as predictors

## References


Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.

Futrell, R., Gibson, E., and Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814.

Gigerenzer, G., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior.* OUP.

Ranjan, S. and von der Malsburg, T. (2024). Work smarter...not harder: Efficient minimization of dependency length in sov languages. In Samuelson, L. K., Frank, S., Toneva, M., Mackey, A., and Hazeltine, E., editors, *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, Rotterdam, Netherlands.

Zeman, D., Nivre, J., et al. (2022). Universal dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.