Dependency length minimization hypothesis revisited

Himanshu Yadav, Shubham Mittal, Samar Husain
hyadav@uni-potsdam.de, shubhammittal1810@gmail.com, samar@hss.iitd.ac.in

The *dependency length minimization (DLM) hypothesis* is widely regarded as a cross-linguistic constraint capturing syntactic complexity in natural languages (Liu et al., 2008; Liu, 2017; Futrell et al., 2015). Although the original formulation in Gibson (1998) uses discourse referents to compute dependency length, a typical method to operationalize dependency length in large scale corpus studies is to count the number of words between syntactically related words. For example, in Figure 1(a), *John* and the verb *met* are separated by 4 words. Since, working-memory is limited, increased dependency length (due to more intervening words) is presumably more complex because it leads to retrieval difficulty during dependency resolution (Gibson, 1998). However, such formulations (e.g., no. of words, no. of discourse referents, etc.) ignore the *syntactic nature* of the linguistic material that intervenes the dependency. For example, in Figure 1(b) *John* and *met* are separated by 4 words, but unlike 1(a), an embedded clause intervenes the dependency. Given the resource constraints, it is reasonable to assume that the complexity of the intervener would lead to greater processing difficulty at the integration site in 1(b). Therefore, the *syntactic* complexity of the intervening linguistic material should be largely minimized (or simplified). We call this the *Intervener complexity minimization (ICM) hypothesis*. Intervener complexity is operationalized as the number of syntactic heads that intervene a dependency. The intervening heads for the dependency *John* $\leftarrow$ *met* in 1 (a) and 1 (b) are underlined in Figure 1; for 1(a) it's 0, for 1(b) it's 1.

In this work, we conduct a cross-linguistic corpus study (involving 52 languages from the Surface-Syntactic Universal Dep. treebanks; Gerdes et al., 2018) to investigate (1) **Independent ICM constraint**: whether the ICM constraint holds independent of the DLM constraint, and (2) **Independent DLM constraint**: whether the DLM constraint holds independent of the ICM constraint. We compare the dependency length/intervener complexity distribution in real language trees with random baseline trees that are similar to real trees in certain properties. In particular, in order to test the Independent ICM constraint, a random baseline was used that matched with the real trees in tree depth, arity, and critically, in dependency length distribution. We call this baseline DL-matched RLAs (see Note 1). On similar lines, the Independent DLM constraint was tested using a random baseline that matched the real trees in tree depth, arity, and critically, in the distribution of intervener complexity. We call this baseline IC-matched RLAs (see Note 1). Linear mixed effects models with varying intercepts and random slope adjustments for languages were used to compare the intervener complexity (for the Independent ICM constraint analysis) and dependency distance (for the Independent DLM constraint analysis). If the Independent ICM constraint is true then the intervener complexity in real language trees should grow slower with sentence length compared to the **rate of growth** in the DL-matched RLAs. Similarly, if the Independent DLM constraint is true then the dependency distance in real trees should grow slower with increase in sentence length compared to the rate in the IC-matched RLAs.

Results from the Independent ICM constraint analysis showed a significant interaction between sentence length and tree type (real vs random) (t=-13.3) such that the average intervener complexity in real language trees grew slower with increase in sentence length compared to the intervener complexity in DL-matched RLAs. However, the Independent DLM constraint analysis showed no significant interaction between sentence length and tree type (t=1.62). See Figure 2.

These results provide compelling evidence that the linguistic material that intervenes syntactically related words tends to be on average simple. It suggests that a complexity measure based on the number of syntactic heads rather than the number of words might be better at capturing the cross-linguistic constraint with regard to syntactic complexity in natural languages. These results also provide evidence for efficiency constraints during language production (MacDonald, 2013).

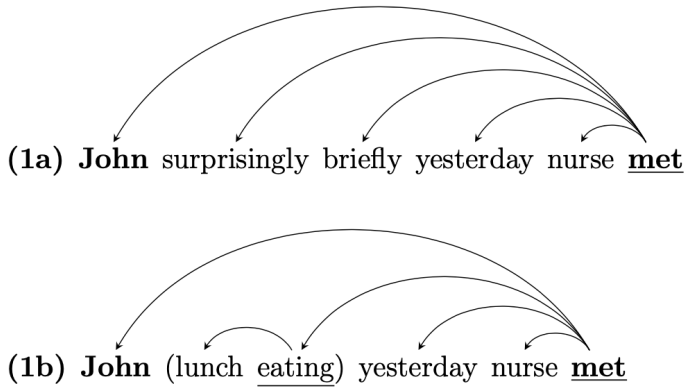

Figure 1: The above examples are glossed translations of Hindi sentences in English. Example 1(a): *John surprisingly met the nurse briefly yesterday.*; Example 1(b): *John met the nurse yesterday after eating lunch.* Dependency arcs go from the head to its dependent.

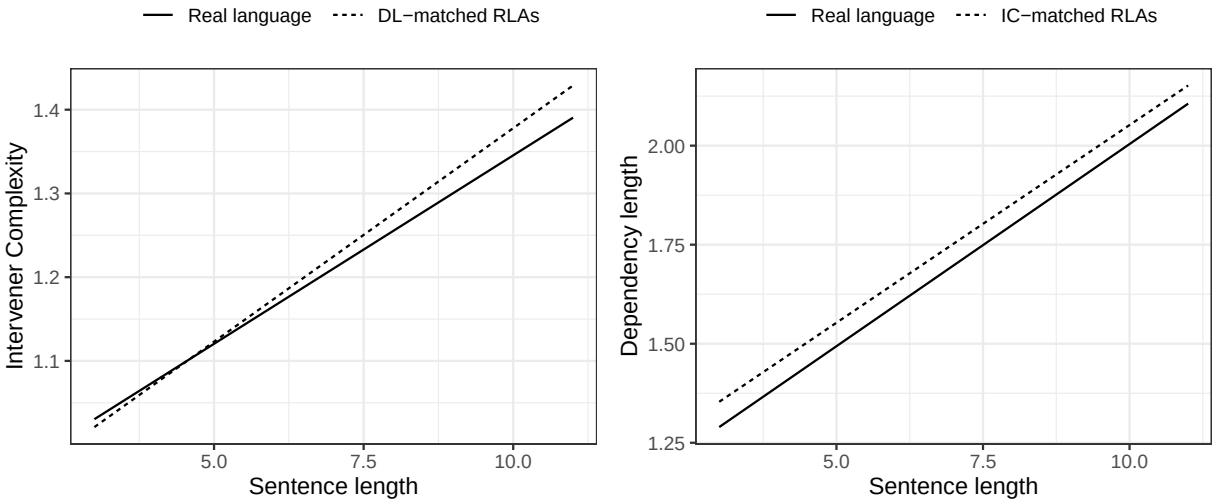

Figure 2: Comparison of Intervener complexity in real trees vs DL-matched RLAs (plot on the left); comparison of Dependency length in real trees vs IC-matached RLAs (plot on the right).

Note 1: Random baselines

DL-matched RLAs: For each real dependency tree, we first generate its random linear arrangements (RLAs) by permuting the linear order of the nodes of the real tree, and then we use rejection sampling to control for dependency length (DL) sequence: the RLAs which do not match in DL-sequence with the real tree are rejected.

IC-matched RLAs: For each real dependency tree, we first generate its RLAs, and then we use rejection sampling to control for intervener complexity (IC) sequence: the RLAs which do not match IC-sequence with the real tree are rejected.

Both the baselines also control for the rate of crossing dependencies using rejection sampling. Since the baseline trees are the RLAs, they match in all the topological properties with the real language trees, e.g., depth, arity, hubbiness, etc. Due to rejection sampling, the tree generation process was prohibitively slow, we have therefore generated trees for sentence length upto 11.

References

Liu, Journal of Cognitive Science (2008); Liu et al., Physics of Life Reviews (2017); Futrell et al., PNAS (2015); Gibson, Cognition (1998); Gerdes et al., Proc. of Universal Dependencies Workshop (2018); MacDonald, Frontiers in Psychology (2013)