Research project: Is the dependency length minimization constraint on human language epiphenomenal?

Himanshu Yadav
Indian Institute of Technology Kanpur
himanshu@iitk.ac.in
https://sites.google.com/site/himanshuyadavjnu/

Introduction

Natural languages possess an important property that they change over time. Languages change in terms of their phonetic inventory, vocabulary, word order choices, among many other things. A common cause of language change, which is often studied by the linguists, is the language contact. When a language comes in contact with another language for sufficiently long period, it can change in its sounds, words, and structural properties, but within certain principles of language change. This approach has told us a lot about what kind of change precedes, what follows, what can change, what cannot, and it can predict possible direction of change for a language. However, this approach fails to answer many questions. For example, how languages would have grown in complexity when they were isolated for centuries? How the human language system as a whole has evolved on this planet? Why does the human language look the way it is today? How we describe the language evolution using a set of simple principles? How do we predict what new structural patterns could emerge in a language, what kind of new languages can evolve, what kind of languages can never exist in human use?

An alternative approach to study language change comes from the complex adaptive system view of language: Languages adapt to the communicative needs and cognitive constraints of its user [1]. The key idea is that the language as a system has emerged from the human mind and it resides in the human mind. It has grown in complexity and keeps growing but bounded by its functional requirements and its environment, the human mind. This approach has the potential to answer many interesting questions about language stated in the previous paragraph.

A well-studied hypothesis under the complex adaptive system perspective is the Dependency Length Minimization (DLM) hypothesis. Under the assumption that languages are shaped by the cognitive constraints of its users, it is hypothesized that languages will prefer structures that minimize working memory load (while maintaining other functional needs). One way to operationalize working memory load is the **dependency length**: the number of words that intervene a dependency[1] (see Figure 1). The longer is the dependency length, the higher is the working memory load [2]. The DLM hypothesis predicts that average dependency length of a sentence from a natural language would be smaller compared to an artificial sentence generated by random reordering of words or by random dependency structures drawn over the same number of words. Specifically, dependency lengths in human language should grow slower with respect to sentence length compared to that in artificial language trees generated under certain topological constraints (that match the human language).[2] The hypothesis is supported by overwhelming empirical evidence from natural languages' corpora [**?**, 3, 4]. The dependency length obtained from natural language trees are significantly smaller (w.r.t. sentence length) than that from artificially-generated baseline trees e.g., from random linear arrangements of language trees. This empirical result holds across languages irrespective of their typology and structural properties.

[1] A hypothetical arc connecting two linguistically related words in a sentence, e.g., a verb and its subject

[2] These artificial language trees can be viewed as a language-like system that has not resided in the human and is free of cognitive constraints like working memory limitation. These could represent a language which evolved outside the human mind, e.g., in machines or in extraterrestrial beings.

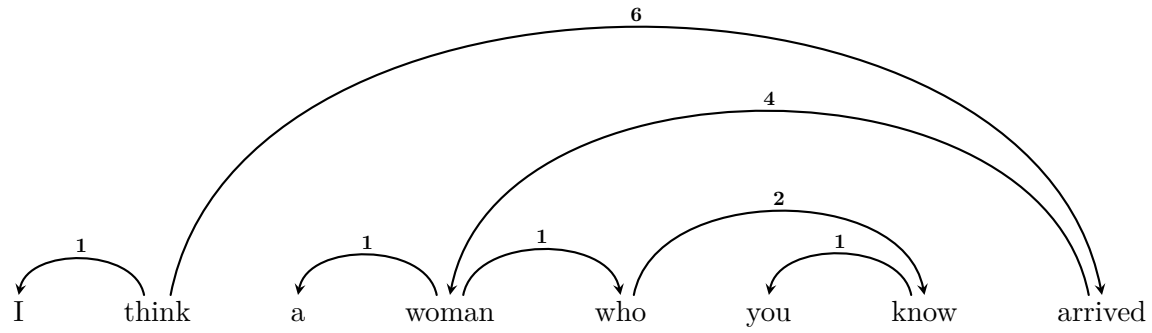

Figure 1

An example dependency tree. Arrows point from heads to dependents. Each dependency arc is labeled with its ***dependency length****: the distance from the head to the dependent, measured in words. The principle of dependency length minimization, which has been proposed as an efficiency-based explanation for syntactic universals of language, holds that dependency lengths should be minimized. In this tree, there are two long dependencies with lengths 6 and 4. English language allows an alternative word order placement here that can shorten the dependency lengths:* ***I think a women arrived who you know****.*

However, DLM is just an approximation of the working memory constraint that influences the language. is it a good approximation? The universality and generalizability of DLM hinges on this question. Is DLM sufficient to characterize the observed effect of working memory limitation in language statistics? Can languages be better characterized by some other measure of memory load? Does the apparent DLM constraint emerges as a consequence of more general constraint on language due to limited memory capacity?

Yadav and colleagues [5] have asked these questions in their corpus study on DLM. The authors propose a new metric called intervener complexity which counts the number of heads intervene a dependency. They claim that the intervener complexity is a better approximation of memory load and a constraint on intervener complexity causes the observed restriction on dependency lengths in human language.

In this project, we aim to test several new measures of working memory load based on the insights from the memory research. For example, feature interference and feature misbinding effects due to memory limitation are commonly observed. We build new graph-theoretic matrices that can approximate working memory load and test which sets of metrics best explain the observed distribution of structural preferences in human language.

Materials and methods

1. Data	Treebanks from the latest version of UD [6]
2. Baselines	Random structures and random linear arrangements controlled for the crossing rate [7]
3. Data analysis	Linear mixed models

Project significance

1. This study will establish standard measures of structural complexity that are best at approximating the working memory constraints on language.
2. The novel measures proposed in this study will be of practical utility in corpus studies and in designing sentence processing experiments.

Project timeline

Jan 2024 – Feb 2024	Determining the measures, writing scripts to extract properties from corpus trees
March 2024 – April 2024	Random baseline generation
May 2024 – June 2024	Data analysis
July 2024 – August 2024	Writing and submission to ACL conferences

Project requirements

The project requires familiarity with dependency syntax, elementary graph theory, linear mixed models and sufficient skills in Python programming.

References

[1] John A Hawkins. A parsing theory of word order universals. *Linguistic inquiry*, 21(2):223–261, 1990.

[2] Haitao Liu, Chunshan Xu, and Junying Liang. Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193, 2017.

[3] David Temperley and Daniel Gildea. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, 4:67–80, 2018.

[4] Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, 2015.

[5] Himanshu Yadav, Shubham Mittal, and Samar Husain. A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, pages 1–22.

[6] Joakim Nivre, Mitchell Abrams, et al. Universal dependencies 2.3, 2018. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[7] Himanshu Yadav, Samar Husain, and Richard Futrell. Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity. *Computational Linguistics*, 48(2):375–401, 2022.