Research Proposal

Project 3.2: Network optimization. Optimization model

Albert Vidal Cáceres & Eva Martín López

30/11/2025

1. Domain and Background

Syntactic dependency structures can be represented as graphs, where each word is a node and each dependency is an edge. These graphs capture the hierarchical organization of sentences and allow the study of linguistic structure using tools from graph theory.

Recent work suggests that the distances in these structures may be close to optimal, not merely shorter than random. According to Ferrer-i-Cancho et al. [1], the dependency distances observed in real sentences can lie very close to the minimum permitted by the tree structure.

In parallel, network theory provides general limits for graph distances, along with methods to compare real networks with both random models and theoretical minima [2]. By integrating this, we can assess whether syntactic networks behave like typical sparse graphs, like optimized structures, or somewhere in between. This makes us question of how the global organization of syntactic graphs positions itself between randomness and optimality.

2. Aims of the research

1. Quantify the topological distance of real syntactic dependency networks and describe how these distances behave across sentences and sentence lengths.
2. Compare these distances with controlled baselines:
   - random graphs with the same number of nodes and edges,
   - graphs that reach the minimum possible mean distance.
3. Evaluate the degree of optimality of syntactic networks by placing them between randomness and theoretical minima, using a normalized measure of distance optimality.

Subgoal for (2): Develop a simple and computationally feasible method to generate the random baseline and to estimate the minimum-distance baseline while keeping the same number of nodes and edges as the real graph. This involves selecting a suitable randomization model and an approximation method for minimal-distance graphs.

3. Hypotheses to be tested

- **H1:** Real syntactic dependency networks have shorter mean distances than random graphs with the same number of nodes and edges, reflecting some structural organization.

- **H2:** The mean distances of real syntactic networks lie closer to the theoretical minimum than to the random baseline, indicating a degree of optimization.
- **H3 (implicit):** If syntactic structures show a consistent degree of optimality, this pattern should appear across sentences of different lengths and not be restricted to specific sentence sizes.

4. Theoretical framework

The project is grounded in graph theory and network science, particularly models of random graphs (e.g., fixed-size random networks) and theoretical bounds for average path-length. This also incorporates recent work on optimal dependency distances and minimal-distance configurations in syntactic networks, including measures that position real graphs between random and optimal configurations. This allows an evaluation of the degree to which syntax adheres to constraints driven by efficiency.

5. Methods

Syntactic dependency graphs will be extracted from the corpus used in previous labs. For each sentence, the mean topological distance of the real graph will be computed using shortest-path distances.

Random graphs with the same number of nodes and edges will be generated to obtain the random baseline.

The minimum-distance baseline will be estimated using theoretical limits for path-length from recent literature.

Finally, an optimality score will be calculated for each graph to quantify where it lies between randomness and theoretical minima. The distributions of real, random, and minimal distances will then be compared across sentences to determine whether syntactic structures exhibit consistent signs of topological optimization.

References

[1] Ramon Ferrer-i Cancho, Carlos Gómez-Rodríguez, Juan Luis Esteban, and Lluís Alemany-Puig. Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308, 2022.

[2] Gorka Zamora-López and Romain Brasselet. Sizing complex networks. *Communications Physics*, 2(144), 2019.