

A Graph Theoretic and Mathematical Framework for the Analysis of Linguistic Structures

Rohit Auxilia E¹, G Jayalalitha²

¹Research Scholar, Department of Mathematics, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India-600117.

Email: rosyfrancis10@gmail.com, <https://orcid.org/0009-0001-2894-8958>

²Professor, Department of Mathematics, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India-600117.

Email: g.jayalalithamaths.sbs@vistas.ac.in, <https://orcid.org/0000-0002-3427-4697>

HOW TO CITE:

Rohit Auxilia E, G Jayalalitha
(2026). A Graph Theoretic and
Mathematical Framework for the
Analysis of Linguistic Structures.
International Journal of Special
Education, 41(16s), 798-810

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

This study proposes a mathematical framework for modelling English linguistic structures using Graph Theory. Various graph models such as Dependency, Tree, Cyclic, Probabilistic, Control Flow, State Transition and Mathematical graphs are used to describe diverse linguistic relationships using selected English Quotes. Mathematical mappings between linguistic elements and graph structures are defined. The study also uses the Principle of Mathematical Induction and the Well Ordering Principle to examine linguistic graph models. Furthermore, the concept of connectivity is explored and appropriate propositions are developed and proven. The proposed framework combines graph theory, discrete mathematics and computational linguistics to provide natural language processing and knowledge representation.

Keywords: Different types of graphs, English Grammar, Mathematical Induction, Well-Ordering Principle, Connectedness, Topological Complexity, Complexity of Topological Order.

1. INTRODUCTION

Graph theory is a subfield of mathematics and computer science that uses diagrams with points and lines to visually describe mathematical concepts. A graph is a visual and mathematical depiction of a collection of things connected by links. Vertices or nodes represent interrelated items, whereas edges, arcs or lines join them together [9]. In computer science, graph theory is the study of mathematical structures that represent pair-wise relationships between objects in a collection. Graphs often known as trees are used to create data structures by connecting vertices and edges together.

The English language has emerged as the main worldwide lingua franca, serving as a means of communication in commerce, education, research and technology [1]. It is the language of international communication, trade and diplomacy with over 1.5 billion speakers worldwide [3]. English's linguistic context is flexible and adaptable. English has acquired words and grammar from several languages as well as developed new vocabulary and expressions. English has become a composite language with various dialects and accents. English's flexibility has made it easier to learn and use in various circumstances.

Natural Language is one of the most complicated forms of human communication, including grammatical, semantic, syntactic and logical links between words and sentences. Representing these linkages quantitatively has emerged as a critical field of study in computational linguistics, Natural Language Processing (NLP), artificial intelligence and knowledge representation. Graph theory is a useful tool for modelling linguistic systems [2]. It uses vertices to represent words and edges to indicate relationships.

Graphs are used extensively in NLP. Syntactic relations connect words in a lexical resource, concepts in a knowledge network and even sentences. These issues are less visible but nonetheless present when addressing types, sentiment analysis, machine translation, document structure and language development [5]. Linguistic graphs differ from linguistic lines, planes and curves which are constructed using linguistic concepts. Linguistic graphs similar to classical graphs consist of linguistic vertices and their connections. Linguistic nodes/ vertices refer to elements from a set of words or terms connected with a linguistic variable [4].

Beyond graph creation, this study looks into mathematical aspects of linguistic graphs. The Principle of Mathematical Induction is used to demonstrate recursive features of graph based linguistic systems whereas the Well-Ordering Principle provides a theoretical foundation for reasoning about ordered linguistic components [6]. Furthermore, the concept of graph connectedness is applied to language graphs to investigate how words and concepts maintain structural and semantic connections within a graph. Several mathematical propositions about graph building, connectivity and language mapping are developed and extensively tested [7].

The proposed framework combines ideas from graph theory, discrete mathematics and computational linguistics into a single mathematical paradigm. By merging graph representations with formal procedures, the work provides both theoretical foundations and computational approaches for evaluating linguistic patterns [8]. This interdisciplinary

technique advances the field of graph-based language modelling and has potential applications in natural language processing, semantic analysis, information retrieval, knowledge graphs, machine learning, educational technologies and AI.

2. PRELIMINARIES

Definition 2.1:

A Dependency Graph is a directed graph that represents the grammatical dependencies between words in a phrase. The mathematical definition is as $DG = (P, Q, R)$ where,

P is the combination of words

$Q \subseteq V \times V$ is the set of directed dependence lines

$R : E \rightarrow D$ labels each edge as subject, object or modifier.

Definition 2.2:

A Tree Graph is a connected, acyclic graph defined by $T(V, E)$ such that

$|E| = |V| - 1$. There is exactly one path between each pair of vertices.

Definition 2.3:

A Cycle Graph is defined as, $CG = (V, E)$ consists of at least one cycle

$$v_1 \rightarrow v_2 \rightarrow \dots \rightarrow v_k = v_1$$

Definition 2.4:

A Probability Graph is represented by $PG = (R, S, P)$ where R is the set of nodes and S is the set of lines and $P: E \rightarrow [0, 1]$. The probability function assigns a probability value between 0 and 1 to each edge in a graph.

Definition 2.5:

A Control Flow Graph is a directed graph, denoted by $CFG = (N, E)$ where, N is the collection of program blocks (or decision nodes) and E is the set of directed edges that represent all possible execution paths between these blocks.

Definition 2.6:

A State Transition Graph is a directed graph, denoted by $STG = (Q, \Sigma, \delta)$ where Q is a finite

collection of states, Σ is a finite set of events or inputs symbols and δ is the transition function that translates each state event pair to the next state.

The graph consists of nodes representing states, directed edges representing transitions and edge labels indicating the event that caused the transition.

Definition 2.7:

A Mathematical graph is an abstract structure that represents interactions between objects. It is defined as an ordered pair $G = (V, E)$ where, V denotes the set of vertices that represent things or entities and E refers to the edges that connect vertices and describe their relationships.

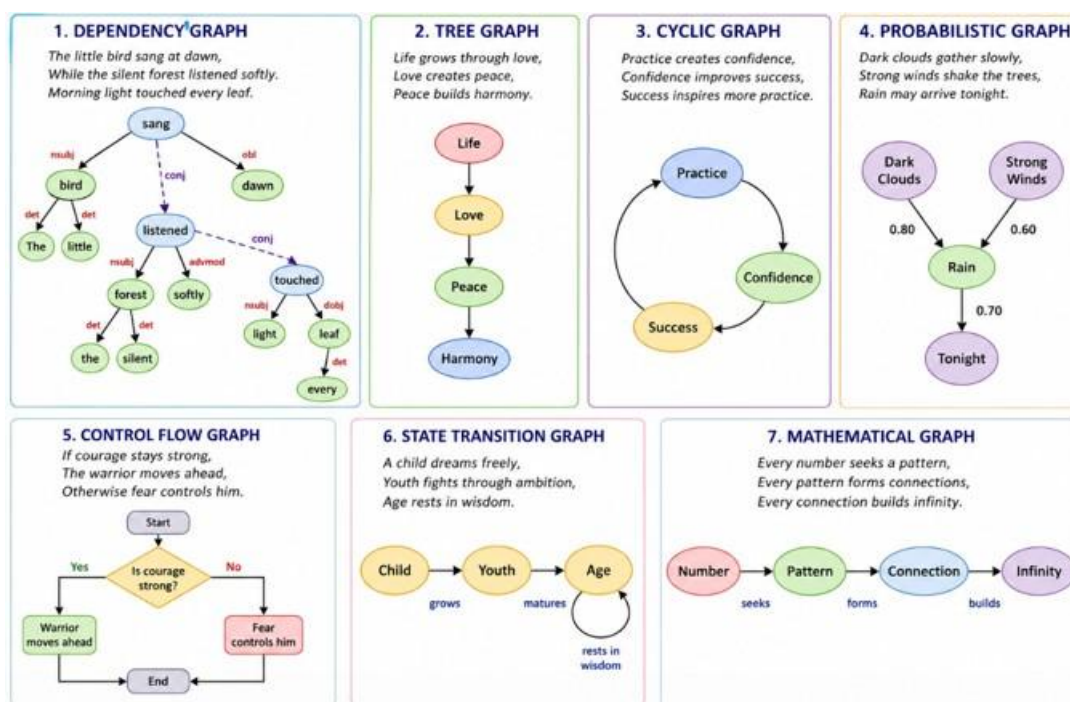


Table 2.1: Quotes for Different types of Graphs

Definition 2.8: The Topological Complexity is a number which measures discontinuity of the process of motion planning in the configuration of space X .

Definition 2.9: The standard time Complexity of a Topological sort is order of $(V+E)$ and the auxiliary space complexity is $O(V)$, where V represents the number of vertices (nodes) and E represents the number of edges.

3. MATHEMATICAL INDUCTION

Let $P(m)$ be a proposition defined for every positive integer $m \geq 1$. If $P(1)$ is true, and whenever $P(r)$ is true, $P(r+1)$ is also true, then $P(m)$ is true for all integers $m \geq 1$.

PROOF: For every positive integer m , $1 + 2 + 3 + \dots + m = \frac{m(m+1)}{2}$

Let $P(m)$ be the statement that $1 + 2 + 3 + \dots + m = \frac{m(m+1)}{2}$. As we have an equation, at each stage of this proof we will show that the left-hand side equals to the right-hand side.

BASE CASE:

When $m = 1$ the LHS = 1 and the right-hand side is $\frac{1(1+1)}{2} = 1$. So, $P(1)$ is correct.

INDUCTION HYPOTHESIS:

Assume that $P(r)$ is true for some positive integer r . That means that the left side of the equation equals the right side, so $1 + 2 + 3 + \dots + r = \frac{r(r+1)}{2}$.

INDUCTION STEP:

We must prove that the statement also holds for $r + 1$.

Consider,

$$1 + 2 + 3 + \dots + r + (r + 1) \\ = \frac{r(r+1)}{2} + (r + 1)$$

1) by the induction hypothesis

$$= \frac{r(r+1)}{2} +$$

$\frac{2(r+1)}{2}$ expressing both terms

$$= \frac{r(r+1)+2(r+1)}{2} \text{ by}$$

combining the fractions

$$= \frac{(r+1)+(r+1)}{2}$$

factor the common term $(r + 1)$

$$= \frac{(r+1)+(r+1)+1}{2}$$

we obtain

So, $P(r + 1)$ is correct. Hence by mathematical induction $P(m)$ is correct for all positive integers m .

3.1 MATHEMATICAL INDUCTION IN LINGUISTICS

Let $P(m)$ be the statement.

A sentence with n adjectives can be grammatically expanded by adding one more adjective.

Using the principle of mathematical induction, we show that a grammatical phrase can include $m \geq 1$.

Base case:

For $m = 1$, consider the following sentence:

“ The cat is small”

This statement has one adjective and is grammatically correct. Therefore, base case is true.

Inductive Hypothesis :

Assume $P(r)$ is true for an integer $r \geq 1$. That is, suppose that a statement with r adjectives is grammatically correct.

For example: The sentence, **“ The cat is small, black, playful, ...”** is assumed to be grammatical.

Inductive Step: To demonstrate that $P(k + 1)$ holds true.

By including additional adjective to the statement containing k adjectives, we obtain a sentence with $k + 1$ adjectives:

“ The cat is small, black, playful, intelligent, ...”

The resulting sentence is grammatically correct due to the recursive and expandable character of adjective sequences in English grammar.

As a result, if $P(k)$ is true, $P(k + 1)$ is also true.

Since: $P(1)$ is true and $P(k) \Rightarrow P(k + 1)$

According to the Principle of Mathematical Induction, $P(n)$ holds true for all integers $n \geq 1$.

Hence, a grammatical sentence can contain any finite number of adjectives.

4. WELL-ORDERING PRINCIPLE

The well-ordering principle states that, every non-empty subset of the natural numbers N has a minimum element. The symbol $\forall S \in N, S \neq \emptyset \Rightarrow \exists m \in S$ such that $m \leq n$ for all $s \in S$.

This indicates that any non-empty collection of natural numbers will always have at least one number that is less than or equal to every other number in the collection.

The well-ordering principle in linguistics defines that every non-empty collection of linguistic structures contains the simplest or least complex component. In linguistics, the Well-Ordering Principle is defined as the concept that linguistic units can be ordered based on a measure of complexity, such as length, morphological structure, or syntactic depth. Under this ordering, every non empty collection of linguistic expressions contains a minimally complex member. This idea is useful in theoretical and formal linguistics, especially in proofs that use minimal counter examples and recursive grammatical structures.

4.1 LINGUISTICS INTERPRETATION

4.1.1 Sentence Length and Complexity:

Let S denote a non-empty set of grammatical phrases. Define a complexity measure c on S such that the word count determines how difficult a phrase is. As word counts are represented by natural numbers, the Well-Ordering Principle ensures that the collection of complexity values associated with S contains at least one element.

As a result, at least one sentence in S has the shortest word count of any sentence in the set. Thus, every non-empty collection of grammatical sentences contains a sentence with a minimal length.

Consider the set,

- “ Birds Fly ” (2 words)
- “ The Birds Fly ” (3 words)
- “ The Small Birds Fly ” (4 words)
- “ The Small Birds Fly Quickly ” (5 words)

The least element according to the word count is, **“ Birds Fly ”**

Thus, the well-ordering principle ensures that any non-empty collection of sentences contains at least one minimally complex sentence.

Word	Complexity
Happy	0
Unhappy	1
Very unhappy	2
Extremely unhappy	3

Thus, **“ happy ”** is the set's smallest member under the specified complexity measure, demonstrating how the Well- Ordering Principle can be used to morphological analyses.

4.1.3. Recursive Clause Embedding:

Let S be a set and $f : S \rightarrow S$ be a function. A recursive definition has the form $a_0 = c$, where c is the base case and $a_n = F(a_{n-1})$, $n \geq 1$, where F is a rule that generates each new element from a defined element.

Morphological Structure and the Well-Ordering Principle:

Let M be a finite set of morphemes where each m_i represents the smallest meaningful unit of language $M = \{ m_1, m_2, m_3, \dots, m_n \}$. A word w can be expressed as an ordered sequence of morphemes such as $w = \{ m_1, m_2, m_3, \dots, m_n \}$. A morphological structure is a mathematical representation of how morphemes connect to form words based on a set of morphological principles.

The number of affixes or changes given to the root form can be used to quantify morphological complexity. In mathematical linguistics, a morphological structure can be represented as a structured collection of morphemes depending on specified formation rules.

For Example: Consider the word family { **happy, unhappy, very unhappy, extremely unhappy** } the lexical item **happy** is the least complex form because it contains no further affixes or modifiers.

Let the number of new morphemes be considered as a measure of complexity.

Recursive clause embedding is a syntactic procedure in which one clause is embedded within another. This process can be continued indefinitely. In other words, a sentence may contain another clause of the same type, resulting in a hierarchical and potentially infinite structure.

In linguistics, recursion is regarded as one of the essential features of human language, since it allows people to produce an

infinite number of sentences from a finite set of patterns.

Example: Consider the sentence, “**John left**”. This is a simple clause and now embedded it with another clause “**Mary thinks John left**”. The clause John left is incorporated within the larger clause Mary thinks John left.

In recursive embedding the process as “**Peter believes Mary thinks John left**” and further “**Susan knows Peter believes Mary thinks John left**”. Each new sentence includes the previous sentence as an embedded clause. Embedding depth is commonly used to measure the complexity of recursive systems.

Sentence	Embedding Depth
John left	1
Mary thinks John left	2
Peter believes Mary thinks John left	3
Susan knows Peter believes Mary thinks John left	4

As the number of embedded clauses increases, so does the embedding depth. The minimally embedded sentence is “**John left**”. The Well-Ordering Principle assures that any non-empty set of such phrases has the least embedding depth.

5. ORDERING AND PARTIAL ORDERING IN LINGUISTICS

Ordering and partial ordering are mathematical concepts used to define the relationships between

words, phrases, sentences and grammatical categories.

Ordering (Total Ordering) in Linguistics:

A total order exists when all elements can be arranged in a linear sequence, with each element having a distinct position relative to the others.

The sentences in a paragraph are presented in a logical order:

ORDER	SENTENCE
S ₁	Time is valuable since it does not return
S ₂	Time should be used properly, because lost time is irreplaceable
S ₃	Every moment is valuable ;consequently, time should be used efficiently
S ₄	Time helps you attain objectives and dreams
S ₅	Time is like a coin; once spent, it is gone.
S ₆	Wasting time requires wasting opportunity
S ₇	Tasks should be prioritized and time managed efficiently
S ₈	Time is a precious resource that should be valued
S ₉	Time is vital to success and fulfilment.

Thus, the idea can be organised as,

S₁, S₂, S₃, S₄, S₅, S₆, S₇, S₈, S₉

This is a comprehensive ordering since each sentence corresponds to a specific position in the arguments sequence.

5.1 PARTIAL ORDERING IN LINGUISTICS:

A partial order emphasizes hierarchical relationships between ideas rather than their linear position. The paragraph includes a central theme, **The Value of Time**. Several supporting ideas are intimately related to this theme, but not

necessarily to one another. A conceptual hierarchy can be expressed as follows:

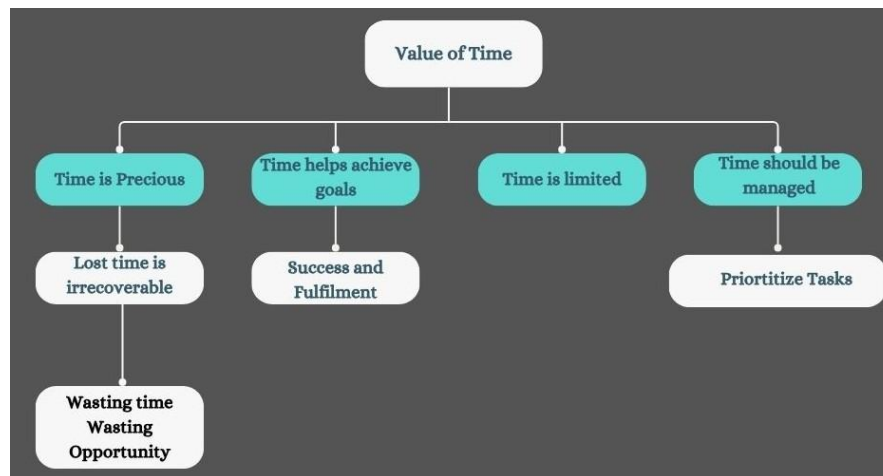


Figure 1: Flow Chart for Value of Time

Partial – Order Relations

Let, A : Value of Time

B : Time is precious

C : Time helps achieve goals

D : Time is limited

E : Time should be managed effectively

F : Lost time is irrecoverable

G : Wasting time means wasting opportunity

H : Success and Fulfilment

I : Prioritize tasks

Then, $A \geq B$, $A \geq C$, $A \geq D$, $A \geq E$

$B \geq F$, $F \geq G$, $C \geq H$, $E \geq I$

However, C and D are not directly equivalent

H and G are not directly equivalent

I and F are not directly equivalent

As a result, this structure represents a partial ordering rather than a total ordering.

Proposition 5.1 : A Linguistic Complexity Bound Based on Local Variations

Let T be a text or discourse, and let $0 < \delta < 1$ be a segmentation parameter.

Let $r = \frac{1}{\delta}$ be the smallest integer which is greater than $\frac{1}{\delta}$. Divide the conversation into m consecutive parts $S_0, S_1, \dots, S_{m-1}$. For each segment S_i , let $V(S_i)$ denote its linguistic diversity as measured primarily by discourse shifts, syntactic complexity, lexical diversity and semantic variance. Let $C_\delta(T)$ denote the overall linguistic complexity of the discourse at scale δ . Then,

$$\delta^{-1} \sum_{i=0}^{m-1} V(S_i) \leq C_\delta(T) \leq 2m + \delta^{-1} \sum_{i=0}^{m-1} V(S_i)$$

EXPLANATION OF THE PROPOSITION:

This proposition establishes a connection between global discourse complexity and local language variance. It asserts that the cumulative variation observed through a discourse's component parts limits the discourse's overall complexity.

The quantity $V(S_i)$ captures the extent to which a segment's language has changed. Such variance could arise from:

Discourse Shifts – which represents changes in topics, viewpoints or communicative functions.

Lexical Diversity – which reflects variations in vocabulary usage

Syntactic Complexity – which reflects variations in sentence structure and clause embedding.

Semantic Variation – which reflects changes in meaning or conceptual content.

Therefore, the proposition formalizes the idea that the accumulation of linguistic diversity throughout a text lead to discourse complexity.

Proof :

Consider a discourse T divided into m consecutive segments.

$$S_0, S_1, \dots, S_{m-1} \quad \dots\dots (1)$$

For each segment S_i , define a linguistic variation measure, $V(S_i)$. It measures the degree of lexical, syntactic, semantic and discourse-level change within that segment.

A segment with more linguistic variance needs a more extensive structural description. As a result, segment S contributes to the descriptive complexity is proportional to

$$\frac{V(S_i)}{\delta} \quad \dots\dots (2)$$

Adding up each segment results in

$$\delta^{-1} \sum_{i=0}^{m-1} V(S_i) \quad \dots\dots (3)$$

which indicates the minimum of information required to account for the observed linguistic diversity across the distance. Therefore,

$$\delta^{-1} \sum_{i=0}^{m-1} V(S_i) \leq C_\delta(T) \quad \dots\dots (4)$$

This gives the lower limit.

To reach an upper bound, consider that conversation fragments are not completely independent. Adjacent segments may have similar language, grammatical patterns, semantic content, or discourse interactions. Such interactions add extra descriptive constraints to the variation recorded within particular segments.

Since neighbouring discourse segments may have overlapping language elements and coherent relationships an additional complexity word must be included. This supplemental contribution is limited to $2m$, where m denotes the total number of segments.

Hence,

$$C_\delta(T) \leq 2m + \delta^{-1} \sum_{i=0}^{m-1} V(S_i) \quad \dots\dots (5)$$

Combining the upper and lower limit yields

$$\delta^{-1} \sum_{i=0}^{m-1} V(S_i) \leq C_\delta(T) \leq 2m + \delta^{-1} \sum_{i=0}^{m-1} V(S_i) \quad \dots\dots (6)$$

As a result, the overall complexity of a discourse is limited by the cumulative linguistic diversity of its constituent segments, as well as a finite correction term that takes into account with inter-segment connections. Thus, language complexity results from the accumulation of local linguistic variances throughout the text.

6. Connected Graph:

Let $G = (V, E)$ be a graph, where V is the set of vertices and E is the set of edges. If there is a path connecting each path of vertices $u, v \in V$, the graph is said to be connected. Formally $u, v \in V \Rightarrow \exists P(u, v)$.

6.1 Linguistic Connectedness:

Let $G = (V, E)$ be a linguistic graph with V a non-empty collection of linguistic units like words, phrases, clauses, sentences or discourse fragments where E is a set of linguistic relations that connect elements of V at the lexical, syntactic, semantic, referential or discourse level.

Linguistically linked graphs (G) have finite paths $u = v_0, v_1, \dots, v_k = v$ such that $(v_{i-1}, v_i) \in E$ for every $i = 1, 2, \dots, k$. In other words, each linguistic unit in the graph is connected to every other linguistic unit by a series of meaningful linguistic links.

Proposition 6.1: Connectedness Preservation in Discourse

Statement : Let $D_n = \{S_1, S_2, \dots, S_n\}$ be a discourse unit. A sentence, clause, paragraph or theme concept may be represented by each discourse unit. Let the corresponding discourse graph be $G_n = (V_n, E_n)$.

A discourse unit S_i is represented by each vertex $v_i \in V_n$, and a significant linguistic relationship between S_i and S_j , such as lexical cohesion, semantic similarity, referential linkage or thematic continuity is represented by each edge $e_{ij} \in E_n$.

Let G_n be a connected discourse graph. The resulting discourse graph $G_{n+1} = (V_{n+1}, E_{n+1})$ stays connected when a new discourse unit S_{n+1} is introduced and incorporated into the existing discourse through at least one linguistic relation with a previously created unit S_k , where $1 \leq k \leq n$.

Proof

Since G_n is connected, there is a path connecting each pair of vertices $v_i, v_j \in V_n$. Let S_{n+1} be the newly introduced discourse unit represented by vertex v_{n+1} . By assumption, S_{n+1} has at least one linguistic relationship with an existing discourse unit S_k . Thus, an edge (v_{n+1}, v_k) is added to the graph.

Because v_k belongs to the connected graph G_n , there is a path from v_k to each vertex in V_n . As a result, for any vertex v_i in V_n there exists a path from v_k to every vertex in V_n . As a result, for each vertex $v_i \in V_n$, a path from v_{n+1} to v_i can be constructed by traversing the edge (v_{n+1}, v_k) and then taking the existing path v_k to v_i .

Hence, v_{n+1} is connected to every other vertex of G_n . Because all vertices in G_n are mutually connected and the newly added vertex is linked to the current graph, every pair of vertices in G_{n+1} is connected by a path. Therefore, G_{n+1} is connected.

7. Quantitative Measures for Linguistic Graphs

Graph-theoretic representations efficiently reflect the structural organisation of English sentences, but they do not provide numerical measurements for evaluating linguistic characteristics. As a result, this section provides five quantitative measurements for linguistic graphs. These measurements make it easier to compare English phrases mathematically and provide information about their complexity, similarity, information richness, recursive structure and grammatical dominance.

7.1 Graph Distance between Sentences

The graph distance between sentences measures the structural difference between two language graphs. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2,$

$E_2)$ denote the two English sentence graphs. The graph distance is defined as $D(G_1, G_2) = |V_1 \Delta V_2| + |E_1 \Delta E_2|$, where Δ represents the symmetric difference between the respective vertex and edge sets. This metric quantifies the structural difference between two English sentences. A smaller graph distance suggests higher structural similarity, while an extended graph distance shows more linguistic complexity and diversity.

Consider the following two English Sentences:

G1 : "Birds fly"

G2 : "The birds fly quickly".

Every sentence is depicted as a linguistic graph with words as the vertices and grammatical relationships between the words as the edges. The linguistic graph G_1 for the first sentence consists of the edge set $E_1 = \{\text{Birds Fly}\}$ and the vertex set $V_1 = \{\text{Birds, Fly}\}$.

Similarly, the vertex set $V_2 = \{\text{The, birds, fly, quickly}\}$ and the edge set $E_2 = \{(\text{The birds}), (\text{Birds Fly}), (\text{Fly quickly})\}$ form the linguistic graph G_2 that corresponds to the second sentence. The symmetric difference operator, represented by the symbol, Δ is used to calculate the graph distance between sentences. The symmetric difference identifies elements that exist in one but not both sets.

In the vertex sets V_1 and V_2 , the common vertices are (Birds Fly), whereas the distinct vertices are (The, quickly). Thus, $|V_1 \Delta V_2| = 2$. Similarly, in E_1 and E_2 , the same edge is (Birds Fly), whereas the unique edges are (The birds) and (Fly quickly).

As a result, $|E_1 \Delta E_2| = 2$. Substituting these values into the graph distance formula

$D(G_1, G_2) = |V_1 \Delta V_2| + |E_1 \Delta E_2|$ such that $D(G_1, G_2) = 2 + 2 = 4$. Therefore, the graph distance between sentences G_1 and G_2 is 4. This finding shows that the second sentence has more lexical and grammatical components than the first sentence, resulting in greater structural complexity.

7.2 Graph Similarity Index

The Graph Similarity Index measures the degree of resemblance between two linguistic graphs. It is defined as

$$S(G_1, G_2) = \frac{|V_1 \cap V_2| + |E_1 \cap E_2|}{|V_1 \cup V_2| + |E_1 \cup E_2|}$$

This metric takes values between 0 and 1, with 1 indicating identical sentence structures and 0 indicating low structural similarity. The graph similarity measure is very effective for plagiarism detection, paraphrase recognition and automated essay evaluation.

Consider the two sentence graphs G_1 and G_2 , where the vertex sets are $V_1 = \{\text{Birds, Fly}\}$ and $V_2 = \{\text{The, birds, fly, quickly}\}$. The common vertices are $\{\text{Birds, Fly}\}$ giving $|V_1 \cap V_2| = 2$.

The union of the vertex sets is $V_1 \cup V_2 = \{\text{The, birds, fly, quickly}\}$ which gives $|V_1 \cup V_2| = 4$. Similarly, the edge sets $E_1 = \{\text{Birds, Fly}\}$ and $E_2 = \{(\text{The birds}), (\text{Birds Fly}), (\text{Fly quickly})\}$. The common edge is $E_1 \cap E_2 = \{(\text{Birds, Fly})\}$. Therefore, $|E_1 \cup E_2| = 3$.

Substituting these values into the similarity formula,

$$\begin{aligned} S(G_1, G_2) &= \frac{|V_1 \cap V_2| + |E_1 \cap E_2|}{|V_1 \cup V_2| + |E_1 \cup E_2|} \\ &= \frac{2+1}{4+3} \\ &= \frac{3}{7} = 0.428 \approx 0.43 \end{aligned}$$

0.43

As a result, the Graph Similarity Index between the two linguistic graphs is 0.43, which indicates that the phrases have approximately 43% structural similarity. A similarity score of 1 shows that the sentence graphs have identical vertex and edge sets, whereas a score of 0 indicates minimal or no structural similarity. Therefore, the Graph Similarity Index is a useful quantitative measure for assessing the structural similarity of English sentence graphs and can be used in linguistic analysis, text similarity, paraphrase detection and natural language processing activities.

7.3 Linguistic Entropy

The concept of linguistic entropy is proposed to measure the degree of ambiguity or information in an English sentence. Claude E. Shannon first introduced the idea of entropy in his seminal work on information theory (1948), defining it as a mathematical measure of information content. Entropy in linguistic graphs quantifies the variety and dispersion of words in a sentence. Higher entropy is typically seen in sentences with a wider range of words, which indicates greater structural complexity and informational richness.

The Linguistic Entropy of a sentence is computed using Shannon's entropy formula is

$$H = -\sum p_i \log_2(p_i)$$

where, H denotes the Linguistic Entropy of the sentence, p_i represents the probability of occurrence of the i^{th} word and n is the total number of distinct words in the sentence. The probability of occurrence of each word is calculated as

$$p_i = \frac{\text{Number of occurrences of the } i^{\text{th}} \text{ word}}{\text{Total number of words in the sentence.}}$$

Shannon's entropy formula uses the logarithm to base two termed as $\log_2$, because information is measured in bits, the standard unit of digital information. A bit (binary digit) represents one or two potential states 0 or 1 and serves as the foundation for the binary number system used in computers. Because current computing systems encode, process and transmit information in binary digits, the amount of data associated with an event is naturally described in bits. Consequently, Shannon chose $\log_2$ to ensure that entropy is measured in binary units. The proposed framework focuses on computational linguistics and graph theoretic analysis; hence the entropy of linguistic graphs is given in bits using $\log_2$.

Example 1: Consider the sentence: “Birds Fly”

Here the sentence contains two words: Birds and Fly. Each word appears exactly once. Therefore, the total number of words is 2 and the probability of occurrence of each word is

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{2}$$

Substituting these vales in Shannon's entropy formula,

$$H = - \left[\frac{1}{2} \log_2 \left(\frac{1}{2} \right) + \frac{1}{2} \log_2 \left(\frac{1}{2} \right) \right]$$

$$\text{Since, } \log_2 \left(\frac{1}{2} \right) = -1$$

the equation becomes $H = - \left[\frac{1}{2} (-1) + \frac{1}{2} (-1) \right] = -(-1) = 1$ bit.

Hence, the linguistic entropy of the sentence "Birds Fly" is 1 bit.

Example 2: Consider the sentences: "The birds fly quickly"

Here each word occurs exactly once. Therefore, the total number of words is 4 and the probability of occurrence of each word is

$$p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$$

Substituting these vales in Shannon's entropy formula,

$$H = - \left[\frac{1}{4} \log_2 \left(\frac{1}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) + \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right]$$

$$\text{Since, } \log_2 \left(\frac{1}{4} \right) = -2$$

the equation becomes $H = - \left[\frac{1}{4} (-2) + \frac{1}{4} (-2) + \frac{1}{4} (-2) + \frac{1}{4} (-2) \right] = -(-2) = 2$ bit.

Hence, the linguistic entropy of the sentence "The Birds Fly Quickly" is 2 bit.

Example 3: Consider the sentences: "The Beautiful Birds Fly Quickly Across Green Fields"

Here each word occurs exactly once. Therefore $p_i = \frac{1}{8}$

Substituting these vales in Shannon's entropy formula,

$$H = - 8 \left[\frac{1}{8} \log_2 \left(\frac{1}{8} \right) \right]$$

$$\text{Since, } \log_2 \left(\frac{1}{8} \right) = -3$$

the equation becomes $H = - 8 \left[\frac{1}{8} (-3) \right] = -(-3) = 3$ bit.

Hence, the linguistic entropy of the sentence "The Beautiful Birds Fly Quickly Across Green Fields" is 3 bit.

Example 4: Consider the sentences: "The Beautiful Birds Fly Quickly Across Green Fields During Early Morning Hours"

Here each word occurs exactly once. Therefore $p_i = \frac{1}{12}$

Substituting these vales in Shannon's entropy formula,

$$H = - 12 \left[\frac{1}{12} \log_2 \left(\frac{1}{12} \right) \right]$$

$$\text{Since, } \log_2 \left(\frac{1}{12} \right) = -3.58496$$

the equation becomes $H = - 12 \left[\frac{1}{12} (-3.58496) \right] = -(-3.58496) = 3.58496$ bit.

Hence, the linguistic entropy of the sentence "The Beautiful Birds Fly Quickly Across Green Fields During Early Morning Hours" is 3.58496 bit.

Table2: Comparative Table for Linguistic Entropy

Sentences	No. of words	Probability of each word (p_i)	Entropy (H)	Information (Bits)
Birds Fly	2	$1/2 = 0.5$	$H = \log_2(2) = 1.000$	1.000 bits
The Birds Fly Quickly	4	$1/4 = 0.25$	$H = \log_2(4) = 2.000$	2.000 bits

The Beautiful Birds Fly Quickly Across Green Fields	8	$1/8 = 0.125$	$H = \log_2(8) = 3.000$	3.000 bits
The Beautiful Birds Fly Quickly Across Green Fields	12	$1/12 = 0.0833$	$H = \log_2(12) = 3.585$	3.585 bits
The Beautiful Birds Fly Quickly Across Green Fields	16	$1/16 = 0.0625$	$H = \log_2(16) = 4.000$	4.000 bits

8. RESULTS

The suggested framework creates a unified mathematical approach for evaluating English Sentences by combining Graph Theory, Topology and Discrete Mathematics. English linguistic structures are represented using Dependency, Tree, Cyclic, Probabilistic, Control Flow, State Transition and Mathematical Graphs. Mathematical Induction reveals the scalability of language structures from simple to complex sentences, but the Well- Ordering Principle and ordering theory provide a formal foundation for sentence structure. The concept of connectivity ensures that sentences and conversation are structurally and semantically coherent. The offered propositions put theoretical limits on language complexity while preserving discourse connection. Furthermore, topological complexity analyses the structural structure of linguistic graphs, whereas topological ordering complexity assesses the dependency-based arrangement of linguistic parts.

Graph Distance measures the structural differences between two language graphs by comparing their vertex set and edge sets. The Graph Similarity Index gives a normalized measure of structural similarity, allowing for objective comparisons between English phrases. Linguistic Entropy quantifies the information content and lexical diversity of sentences, with higher entropy levels suggesting greater linguistic complexity. In the sentence the words increase and the bits also increases.

Entropy increases as the number of distinct words in a sentence grows, assuming each word is equally likely to appear. This reflects greater information content and linguistic complexity. Consequently, longer sentences with larger vocabularies tend to have higher entropy values than shorter ones, making entropy a valuable quantitative measure of information richness in linguistic graph analysis which is shown in the table 1. Overall, the framework establishes a strong mathematical foundation for modelling, evaluating and validating English phrases for use in computational linguistics, natural processing and knowledge representation.

9. CONCLUSION

This work shows that fundamental mathematical concepts can be successfully used to the analysis of linguistics structures, including mathematical induction, the well ordering principle, ordering and partial ordering, linguistic complexity bounds and graph theoretic connectedness. The well-ordering principle proves the existence of minimally complicated language forms, whereas mathematical induction offers a framework for simulating the recursive evolution of linguistic structures. The hierarchical and dependent relationships between language units at the morphological, syntactic, semantic and discourse levels are described by ordering and partial ordering.

The linguistic complexity proposition demonstrates how local variations, such as lexical diversity, syntactic complexity, semantic variation and discourse shifts accumulate to produce global

discourse complexity. Furthermore, by depicting conversation as network of interconnected units, linguistic connectedness provides a mathematical characterisation of coherence. According to the proposition on connectedness maintenance in discourse, coherence is preserved when newly added discourse units continue to be related to existing units through significant linguistic relationships. When taken as a whole, ideas offer

a cohesive mathematical framework for comprehending linguistic complexity, discourse organization and language structure. In addition to creating new avenues for study in mathematical linguistics, discourse analysis, graph theory and fractal language studies, the integration of these theories enhances the relationship between mathematics and linguistics

REFERENCES

- Nastase, V., Mihalcea, R., & Radev, D. R. (2015). A survey of graphs in natural language processing. *Natural Language Engineering*, 21(5), 665–698. <https://doi.org/10.1017/S1351324915000340>.
- Vasanth, W. B., Ilanthenral, K., & Smarandache, F. (2022). *Linguistic Graphs and their Applications*. <http://www.gallup.unm.edu/eBooks-otherformats.html>
- Karmakar, S. (n.d.). *Rooted Acyclic Directed Graphs and Indian Grammatical Tradition: A Lexicalist Framework*. <https://ssrn.com/abstract=4253163>
- Piperski, A. (2014). An application of graph theory to linguistic complexity. *Yearbook of the Poznan Linguistic Meeting*, 1(1), 89–102. <https://doi.org/10.1515/yplm-2015-0005>
- Bondy, J. A., & Murty, U. S. R. (n.d.). *GRAPH THEORY WITH APPLICATIONS NORTH-HOLLAND New York • Amsterdam • Oxford*.
- A *Graph-Theoretic Framework for Capturing Structural Similarities Among World Languages*. (n.d.). <https://ssrn.com/abstract=6715505>
- Kenneth Falconer, *Fractal Geometry Mathematical Foundations and Applications*. 3rd edition, John Wiley & Sons limited, England.
- Kenneth H. Rosen, *Discrete Mathematics and its Applications*, 8th edition, MC Vraw Hill education, New York 2019.
- Chandrakant R. Gujar, *Graph Theory and its Applications in Computer Science and Technology* published by JETIR. ORG January 2023, Volume 10 Issue 1 www.jetir.org (ISSN-2349-5162).