Compositionality in Large Language Models: Testing VP-Ellipsis with BERT

Yexiang (Tom) Tang

Department of Philosophy

Washington University in St. Louis

Honors Study

Dr. Nick Danis, Dr. Matt Barros, Dr. Kit Wellman

April 15, 2025

Abstract

By testing with Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), this paper investigates whether Large Language Models (LLMs) exhibit compositionality. Compositionality, the principle that complex meanings are systematically built from components and their structure, is central to philosophy, linguistics, and cognitive science. Traditional theories like Jerry Fodor's Language of Thought Hypothesis (LoTH) conceptualize cognition in terms of rule-governed symbolic operations. In contrast, LLMs use distributed representations. In addition to the success of LLMs in recent years, there are debates on whether LLMs' success is due to the true capture of compositionality or merely mimicking it. To address this issue, this paper tests BERT's embedding vectors by investigating whether the model detects silenced structures in elliptical sentences. If BERT encodes the silenced structure with measurable similarity to their explicitly stated counterparts, it would be a sign for compositional understanding. Based on the measurements of cosine similarity and Euclidean distance between embedding vectors, the findings suggest that BERT does exhibit signs of compositionality but in an inconsistent way. These results add to the broader debate on LLMs' compositionality and indicate the need for further study into the analysis of LLMs' natural language processing.

Keywords: Philosophy of Mind, Functionalism, Language of Thought, Large Language Models

Compositionality in Large Language Models: Testing VP-Ellipsis with BERT

Compositionality—the principle that complex meanings derive from the arrangement of simpler parts—has long been at the heart of linguistic and cognitive theories. Traditional, rule-based frameworks like Jerry Fodor's Language of Thought Hypothesis (LoTH) posit that cognition is grounded in a structured "mentalese," with symbolic operations over discrete concepts (Fodor, 1975; Fodor, 2008). By contrast, large-language-models (LLMs), driven by pattern-learning algorithms and distributed representations, raise the question of whether their remarkable performance truly reflects compositionality. In the following sections, we first introduce the philosophical and computational background that motivates this debate, then present an experiment on BERT's ability to detect the silenced structure in elliptical sentences—a key indicator of compositional processing.

Background

Philosophy and Linguistics Background

Representationalism and Language of Thought Hypothesis (LOTH)

The ongoing debate of whether LLMs encodes compositionality fits in a broader philosophical debate: *Can we duplicate human intelligence in machines?* Compositionality—the principle that the meaning of an expression derives from its components and structural arrangement (Chomsky, 2002; Fodor, 2008; Hupkes, 2020; Partee, 2001)—is particularly significant here because it is required by a fundamental assumption about human mind: the mind operates as a compositional system over symbolic representations (Fodor & Pylyshyn, 1988; Rescorla, 2023). This section breaks down this idea and starts with the representational theory of mind. Representationalism argues that intentionality, the aboutness of mental states, involves representations of states of affairs (Lycan, 2023). That is to say, whenever we think, feel, sense, or believe, our mind is representing objects or scenarios that can be real or unreal. Like a map depicts the layout of a town, the internal mental representation in our mind refers to a situation in the external world. Moreover, the representational theory is also applied to be a theory of

consciousness, introspection, and the subjective qualities in experiences (Gennaro, 2008; Kriegel, 2009; Tye, 1997). There are variances within the representational theory itself such as the division between strong version vs. the weak version (Block, 1996; Drestke, 1993; Tye, 2003). For the purpose of this paper, the basic outline of the theory above is sufficient.

Based on representational theory, the Language of Thought Hypothesis (LoTH), by Jerry Fodor (1975, 2008), argues that our thinking is the operation of representations under certain rules. Like our language use runs with a set of rules, our thoughts take place in a formal mental language, known as *Mentalese*. Not surprisingly, Mentalese shares multiple key features with natural languages, such as the use of discrete concepts, semantic properties, and compositionality. A widely accepted definition of compositionality is that the meaning of a complex expression is determined by the meanings and arrangement of its simpler components. Compositionality is considered fundamental to our cognitive and linguistic abilities, and it enables us to: 1. Understand new expressions using familiar components (Fodor & Pylyshyn, 1988), 2. recombine known elements into new expressions (Chomsky, 1956), and 3. substitute words with synonyms without altering the meaning of an expression (Pagin, 2003).

In recent decades, LoTH has found extended applications in cognitive science, where it is seen as a robust framework for understanding human cognition (Kazanina, 2023; Kemp et al., 2008). Some empirical studies have demonstrated that structured symbolic representations, as proposed by LoTH, are crucial in various cognitive tasks, including object tracking, reasoning, and concept acquisition in both humans and non-human animals (Quilty-Dunn, 2022). Alongside these empirical investigations, cognitive scientists also aim at the theoretical construction of fundamental human cognition by connecting to LoTH.

Functional-Computational-LoTH Framework of Mind

As a theory of mind, LoTH is usually incorporated with a computational and functional theory and comes together with a comprehensive framework of mind (Fodor, 1990; Rescorla, 2023). Functionalism defines mental states as their functional roles in mental activity (Levin, 2023), while computationalism takes mental activities as computations over symbolic

representations (Rescorla, 2020). Here, I refer to this theoretical frame as the *Functional-Computational-LoTH framework*,[1] which gives a theory that the human mind is a computational system that operates over structured representations, where operations are carried out according to the rule of *Mentalese*. Much like a Turing style finite-state machine with each of its operations given in a command list, mental activities take place according to systematic, rule-governed manipulations over symbolic representations.

Though not aimed at a theory of language, the Functional-Computational-LoTH framework nonetheless shed light on linguistic questions, as Fodor (1975) admitted that Mentalese and natural language mutually provide supporting evidence. Interestingly, LoTH framework aligns with Chomsky's theory of syntax in its reliance on rule-based operations on discrete symbols (Chomsky, 1956). According to Chomsky's phrase structure grammar, sentences are built from smaller constituents, i.e. words and morphemes, much in the same way that thoughts in LoTH are constructed based on symbolic representations in Mentalese.

Distributional Semantics

Chomsky (2002) famously argued that finite-state machines are inadequate for capturing natural language syntax, proposing instead a rule-based, context-free grammar that became central to generative linguistics. In contrast, the rise of Large Language Models (LLMs) has renewed interest in data-driven approaches, prompting fresh discussions in both philosophy of mind and linguistics (Lew et al., 2020; Mahowald et al., 2024; Peterson, 2023). One of the most striking shifts is the framework of *distributional semantics*, which differs fundamentally from traditional, symbolic models. Instead of relying on pre-defined component units and symbolic rules, distributional semantics assumes that word meanings are derived from patterns of contextual usage. This approach represents words as points in a high-dimensional space, where proximity corresponds to similarity in context (see Figure 1.1 & 1.2; Clark, 2015; Erk, 2012). By

[1] The term Functional-Computational-LoTH framework is not standard in the literature. However, it is introduced here to succinctly capture a widely accepted conceptual alignment among functionalism, computationalism, and the Language of Thought Hypothesis (LoTH) in explaining mental processes.

transforming written texts into numerical vectors, distributional semantics allows machines to process and compare words based on co-occurrence patterns, enabling the analysis of vast amounts of text data. This departure from categorical distinctions marks a significant shift in how linguistic meaning is conceptualized and operationalized in computational systems.

Figure 1.1

An example of embedding segments of words into vectors.

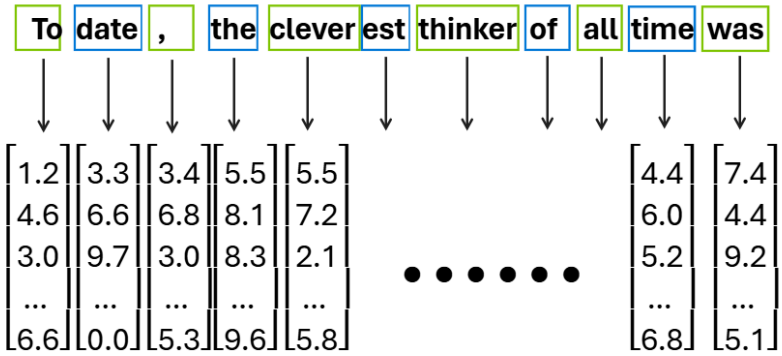
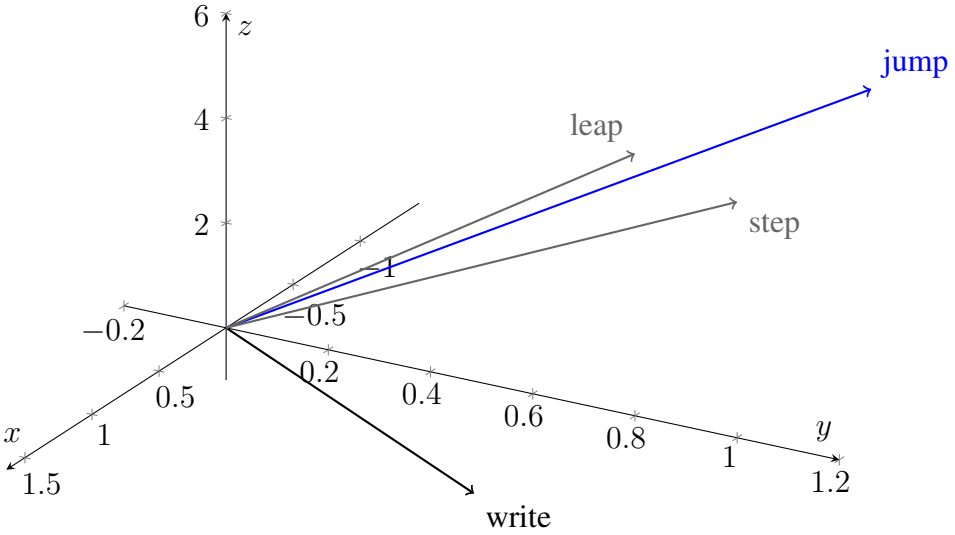


Figure 1.2

Similar words have closer vector representations in high-dimensional space



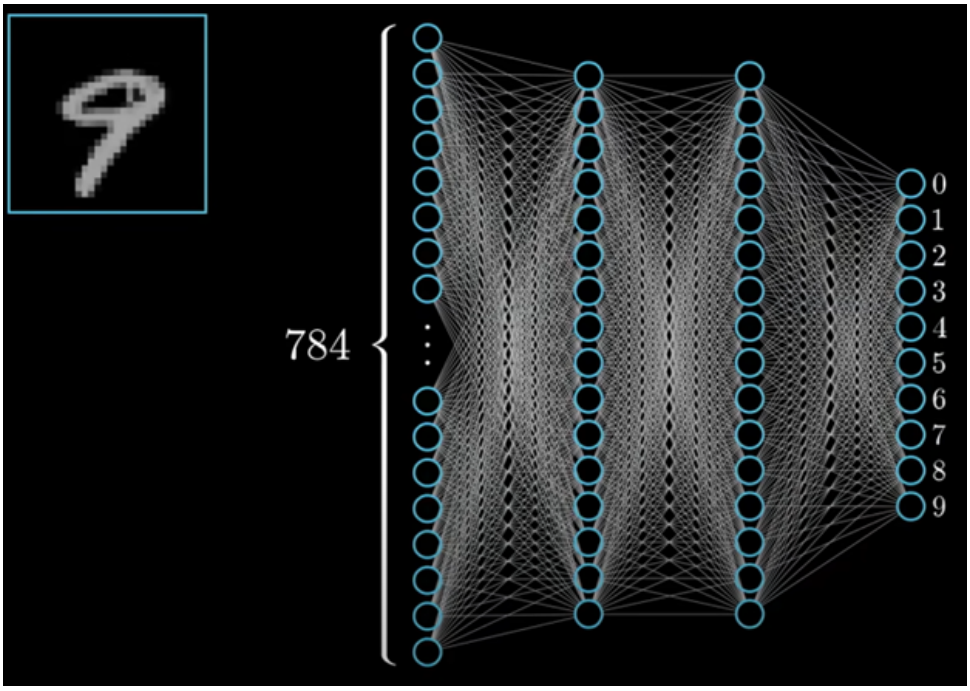
Computer Science Background

Artificial Neural Networks (ANNs)

Beyond the theoretical contribution from distributional semantics, the recent empirical success of Artificial Neural Networks (ANNs) models has brought new debates on both human mind and natural language. ANNs, developed based on perceptron models first introduced by Rosenblatt (1958), consist of multiple layers of interconnected units inspired by-but critically different from-the structure of brain neurons. Each unit receives value from the units in the preceding layer, processes them, and passes the results to the next layer. A classic example of an ANN application is recognizing handwritten numbers. Suppose we have a 26×26 pixel image, which consists of 784 pixels in total. In the ANN, each of these 784 pixels is assigned to an input unit, where its value corresponds to the intensity of the pixel. Multiple hidden layers process these inputs, and the network concludes with 10 output units, each representing one of the possible digits (0 through 9, see Figure 1.3). This example is chosen for its simplicity and clarity in demonstrating the basic operations of an ANN. Image recognition tasks, such as digit classification, involve straightforward, visual inputs that are easier to conceptualize and explain compared to the high-dimensional, contextual embeddings used in natural language processing (NLP). By starting with image recognition, we provide a foundational understanding of how an ANN processes input data, builds representations, and makes predictions. This foundation is crucial for appreciating the more complex mechanisms at work in NLP applications, which will be addressed later.

Figure 1.3

An example of ANN recognizing hand-written "9"

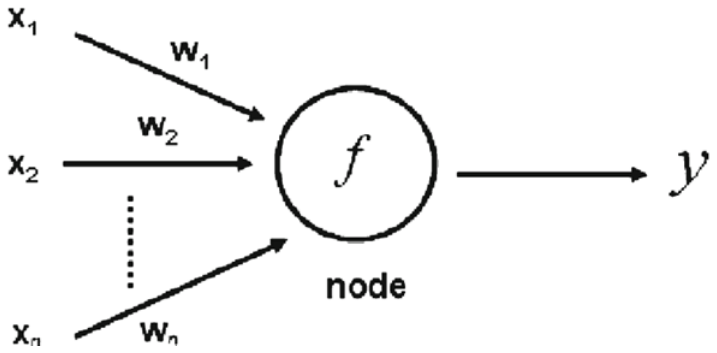


Note. Except input layer, each unit is connected to units in the previous layer. Retrieved from 3Blue1Brown.

For each connection between two nodes, a weight is assigned that multiplies the input from the preceding node. A given unit activates if the weighted sum of its inputs exceeds a designated threshold. The unit then applies an activation function to this sum, producing an output which is passed to the next layer (see Figure 1.4; Zou, 2008).

Figure 1.4

A sketch of input and output for a single neuron in ANN

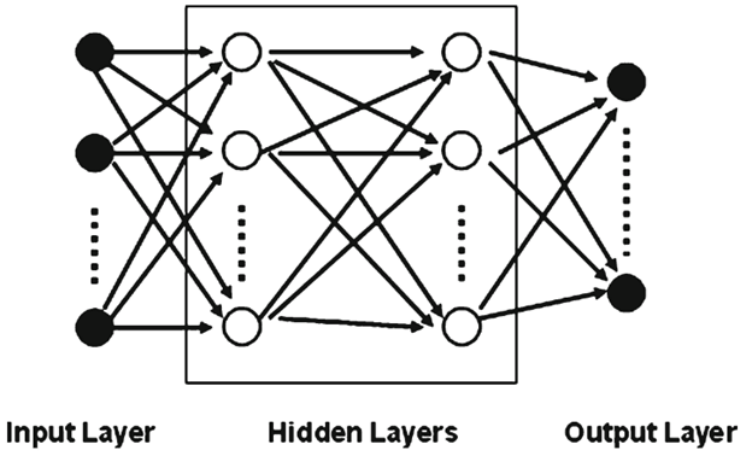


Note. $\{x_1, x_2, \ldots, x_i\}$ *are inputs;* $\{w_1, w_2, \ldots, w_i\}$ *are weights;* f *is the function in the neruon;* y *is the output.*

This process is repeated across multiple layers, where each connection between neurons has different weights and biases. The weights determine the strength of the connection between neurons, while the biases act as adjustable thresholds that help the network shift the activation of neurons independently of the input values. Together, weights and biases allow the network to flexibly adjust its computations to better fit the training data. The initial layer, which receives the raw input data, is called the input layer, while the final layer producing the prediction is the output layer. The intermediate layers are known as hidden layers (see Figure 1.6; Zou, 2008). Through a process of iterative adjustment, known as back propagation training, the network "learns" by modifying the weights and biases to minimize errors in its predictions.

Figure 1.5

A sketch of overall ANN layer structure



Note. Retrived from Zou, 2008

The parameter adjustments depend on the requirements of the specific task. And one of the most fascinating aspects of this learning process is its potential to provide a theoretical framework for understanding how language acquisition occurs (Rumelhart & McClelland, 1986). Indeed there has been some debate among linguists about ANNs' contribution to learning theories across various domains, including phonology (Heinz, 2016) and syntax (Clark, 2017).

LLMs and Transformer Model

The developed ANN-based architecture varies depending on the input size, type, and purpose of the task (Corssberg, 2013; Rio-Torto, 2020). Among these architectures, Large Language Models (LLMs) have recently gained significant attention across multiple fields (Mikolov et al., 2013). Notable LLMs include ChatGPT, Llama, and PaLM, which all are specifically designed for Natural Language Processing (NLP) tasks, such as automatic translation, text summarization, and text generation. And the key innovation for their success is the implementation of the *Transformer* model, which involves two algorithms: *word embedding* and *attention layers*.

Word Embedding. Rooted in the principles of distributional semantics, word embeddings convert words into numerical representations. Initially, these representations are assigned

randomly, but through iterative training and parameter updates, the embeddings evolve to capture semantic relationships. Words that frequently appear in similar contexts, such as "king" and "queen" or "dog" and "cat," are assigned similar vector representations in a high-dimensional space. This vector-based approach allows LLMs to effectively perform NLP tasks by encoding and preserving semantic relationships between words.

However, natural language is more complex than isolated word meanings; the context in which words are used is essential for accurate understanding, especially in long texts. Basic word embeddings, while powerful, are insufficient for capturing this level of nuance. This is where the *attention* mechanism comes into play.

Attention mechanism. Transformers enhance traditional ANNs by incorporating multiple attention layers, which calculate the relationships between all words in a text. By assigning varying levels of "attention" to words based on their relevance, the model captures both the meaning of individual words and the broader structure of the text. For instance, the word "model" may have different meanings in "Large Language Model" and "fashion model." The attention mechanism helps the model discern such differences by analyzing the relationships between "model" and the surrounding words.

Additionally, Transformers utilize multi-head attention, which evaluates various aspects of word relationships, such as syntactic and semantic dependencies, simultaneously. Unlike traditional approaches that treat syntactical and semantic dependencies as discrete functional categories, Transformers encode these relationships in terms of correlations within a high-dimensional space. By doing so, the model captures nuanced connections between words, allowing it to reflect connections among words as continuous gradients of influence rather than predefined rules. After computing these relationships, the attention mechanism updates the vector representation of each word to reflect its contextual meaning. The ability to process and contextualize these relationships makes Transformers particularly efficient for NLP tasks (Gillioz, 2020).

The Focus of Discussion: Compositionality

In contrast to the Functional-Computational-LoTH framework, LLMs adopt a fundamentally different approach to process natural language. Rather than relying on a set of symbolic, rule-based operations, LLMs process data by adjusting parameters of real numbers based on the input they receive. This adaptive process is grounded in mathematical computations within each unit, based on corresponding weights and biases. Whenever LLMs make predictions or classifications, they compare their outputs to the expected results. The difference between the actual and expected outputs is measured with some error metric, which quantifies how far the model's performance deviates from the desired outcome. To minimize this error, LLMs rely on *backpropagation*, a method that minimizes the error by adjusting the weights and biases in each layer. These adjustments are guided by *gradient descent*, a technique commonly used in the process of error-minimization. Through numerous iterations of this process, the model gradually "learns" how to get tasks done well.

This process of learning is very different from the model over symbolic representations, such as LoTH. Symbolic representation machines rely on explicitly defined and interpretable operations, while LLMs adapt dynamically based on their performance during training. This leads to a sharp distinction between the two approaches of intelligence simulation: LoTH takes the mind as a structured, rule-based system operating on discrete symbols, whereas LLMs operate on adaptive learning principles, using continuous computations and correlations to process data. And LLMs' success in a wide range of applications further emphasizes this distinction.

To resolve this puzzle, a number of researchers focus on finding effects of compositionality in LLMs. As defined, compositionality is the property that complex representations are constructed from simpler parts, and the meaning of a complex representation is determined by the meanings of its components and the way they are combined. For example, the meaning of the sentence "John loves Mary" is determined by the constituent words "John," "loves," and "Mary" and the way they are arranged. Notably, even though the sentence "Mary loves John" contains the same words, its meaning differs because their arrangement. Aiming to

bridge the gap between LoTH and ANNs, researchers are focused on this particular question: do LLMs demonstrate compositionality?

There is much support for both sides on this question. Supporters of the 'NO' side typically refer to three points: 1) there is no use of discrete concepts in LLMs in general. As the models do not rely on concepts for operations, they cannot demonstrate compositionality in human sense. For instance, it is unimaginable to construct the meaning of the sentence "the cat sat on the mat" without having the concept of "cat" "sat" or "mat" at all (Biever et al., 2023; Bishop, 2021; Lopes et al., 2023; Nefdt et al., 2022). 2) LLMs are solely trained on output prediction without any understanding of semantic meaning (Bender, 2020). 3) LLMs fail in tests particularly designed to test compositionality (Baron, 2019; Lake, 2018).

On the other hand, proponents of the 'yes' camp are focused on the following points: 1) LLMs succeed in recognizing symbolic structures (Bowman et al., 2015; Locatello et al., 2020; Potts et al., 2015). 2) LLMs capture the essential parts of linguistic meaning (Liang et al., 2015; Pavlick et al., 2023). 3) LLMs succeed in handling new tasks, suggesting that their operations are built from part–a sign of compositionality (Tamir et al., 2023). Interestingly, it's also been found that while LLMs pass some tests for compositionality, they fail others, and performance varies across different models (Hupkes et al., 2020).

A key challenge in the debate on compositionality in LLMs is the inconsistent use of terms like "semantic meaning" and the variety of methods employed by researchers. In this paper, meaning is defined as a truth-conditional function—where a sentence's meaning is determined by the conditions under which it is true or false (Davidson, 1967). This definition is compatible with computational contexts, for both human mind and machines, avoiding debates over which definition of meaning is best.

Due to the variety of methods and models current researchers test with, from recurrent neural networks (RNNs) to tree-structured models and Transformers, the debate is still ongoing. This paper focuses on Transformer-based models because their architecture offers a strong framework for addressing these differences. The rationale for this choice and the methods used

will be explained in the *Method* section.

Method[2]

This paper investigates compositionality by examining the embedding vectors within LLMs. The experiment analyzes how **semantic** and **structural** similarities influence these vectors, focusing on sentences that vary in surface structure and meaning. Specifically, we evaluate whether Bidirectional Encoder Representations from Transformers (**BERT**) captures similarity between sentences that differ in surface form but share hidden structural features grounded in linguistic theory. According to the property of compositionality, the meaning of a complex representation should derive from the meanings of its components and structures. This analysis helps us assess whether BERT demonstrates compositionality in its embeddings.

Model Selection: BERT

The experiment uses the *bert-large-uncased*[3] model introduced by Devlin et al. (2019). BERT was chosen for several reasons. Firstly, BERT is an advanced model for generating contextual word embeddings. Its transformer-based architecture uses self-attention mechanisms to handle long-range dependencies and relationships within sequences. This mechanism enables the machine to consider contexts of words. Unlike unidirectional models, BERT's bidirectional training considers both left and right contexts simultaneously. This significantly enhances the machine's performance on tasks that require a comprehensive grasp of word relations and sentence structure.

Specifically, BERT takes a raw text sentence as input and returns a high-dimensional embedding vector as output. This design is critical for the experiment, as it enables direct analysis of the models internal representations without the interference of text-generation objectives, which are common in more mature language models like ChatGPT. By focusing on these

[2] For codes, dataset, and visualizations of results, please refer to the Github repository https://github.com/yexiang-tang/VPEinBERT

[3] The particular version of BERT model used is the Whole Word Masking Cased variant of BERT-Large. https://github.com/google-research/bert

embedding vectors, the experiment explores how the model processes and represents text at a fundamental level.

Case Study: Ellipsis

Ellipsis is a linguistic phenomenon in which certain parts of a sentence are omitted but remain understood based on context. Consider this example:

(1) *Sarah teaches history, but John does not.*

(2) *Sarah teaches history, but John does not teach history.*

In (1), the verb phrase “teach history” is not pronounced in the second clause but still understood, making (1) an instance of **Verb-Phrase Ellipsis** (VP-Ellipsis). Under **the structural approach** to ellipsis (Merchant, 2018), which this paper takes, ellipsis contains unpronounced syntactic parts that must be recovered based on sentence structures. This contrasts with purely semantic or discourse-based theories, which assume that ellipsis is resolved through contextual inference alone.

Compositionality states that the meaning of a complex expression is determined by the meanings of its parts and their structure. A truly compositional model should represent both the overt and omitted parts. In (1), "John does not" is incomplete on its own but is easily understood as "John does not (teach history)" based on sentence structure. If a model correctly reconstructs this omitted part, it suggests that the model tracks syntactic dependencies rather than merely recognizing word patterns. This is important because compositionality is not about memorizing phrases but about systematically deriving meaning from how words and structures come together. Successfully detecting ellipsis means the model encodes syntactic structure in a way that is similar to human compositional reasoning.

To investigate whether BERT detects the silenced structure, we compare how it represents (1) and its fully stated counterpart (2). If BERT’s embeddings show strong similarity between (1) and (2), while distinguishing them from unrelated contrasts, this suggests that it detects the silenced syntactic structure. By focusing on embedding vectors rather than raw text output, this

approach directly tests BERT's internal structure, which excludes the compounding effects of fine-tuning.

Dataset

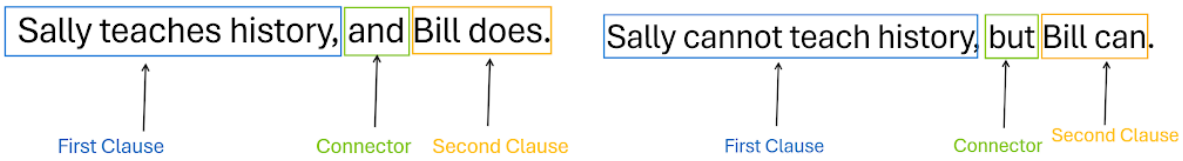
The dataset consists of systematically generated sentence triplets where each triplet contains the following sentences:

1. **Full Sentence**: Complete sentences with explicit syntactic and semantic content.
    - Example: *"Sally teaches history, and John teaches history."*
2. **Elliptical Sentence**: Sentences where a verb phrase is omitted.
    - Example: *"Sally teaches history, and John does."*
    - This tests whether BERT encodes the hidden structure necessary for ellipsis understanding.
3. **Contrast Sentence**: Sentences that have identical surface form with elliptical sentences but have different semantic meanings.
    - Example: *"Sally teaches history, and John hikes."*
    - This serves as a control to determine whether BERT truly captures underlying syntactic relations or merely relies on surface-level similarities.

Each elliptical sentence follows a two-clause structure, where the clauses vary in syntax, verb tense, and auxiliary verbs. The coordinating conjunctions "and" or "but" are used depending on the logical relationship between clauses. For example, in *"Sally teaches history, but John does not,"* the first clause (*"Sally teaches history"*) is classified as **Positive Present**, while the second clause (*"John does not"*) is **Negative Present**. Similarly, in *"Sally cannot teach history, and John can,"* the first clause (*"Sally cannot teach history"*) is classified as **Negative Auxiliary**, while the second clause (*"John can"*) is **Positive Auxiliary**.

Figure 2.1

Two sample elliptical sentences



Differing in syntax, verb tense, and auxiliary verbs, sentences in dataset can be categorized into three main syntactic types, as illustrated below:

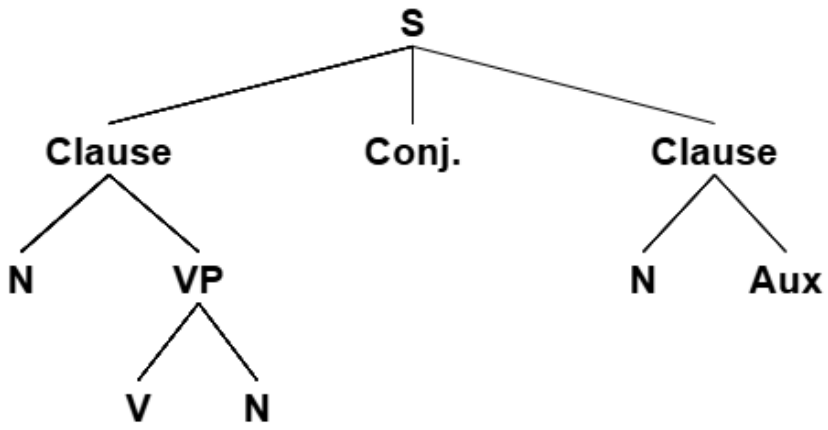
1. **Basic Clause Structure (type-0)**:

The first clause contains a subject noun (*N*), a verb phrase (*VP*), and an object noun. And the second clause consists of only a subject noun and an auxiliary verb (*Aux*), reflecting ellipsis.

Example: *"Sally teaches history, and Bill does."*

Figure 2.2

The syntactic tree of basic clause structure

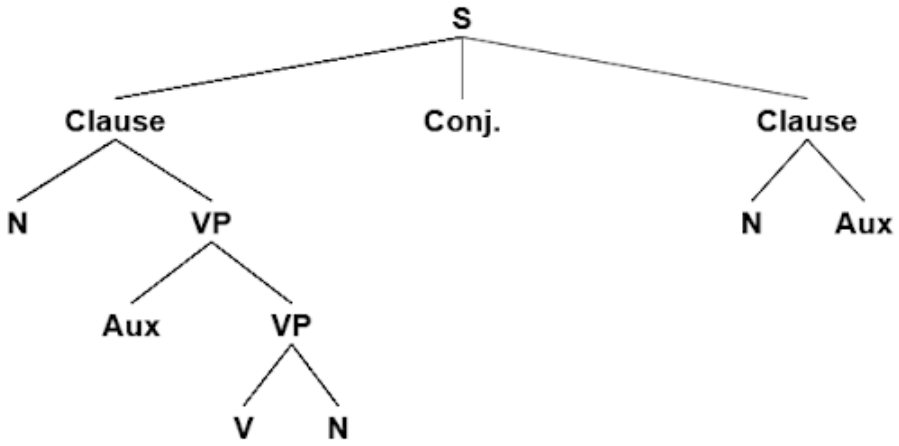


2. **Auxiliary-Enhanced Clause Structure (type-1):**

The first clause contains an auxiliary verb (*Aux*) before the main verb phrase (*VP*). The second clause consists of a noun and an auxiliary verb. Example: *"Sally can teach history, but Bill cannot."*

Figure 2.3

The syntactic tree of auxiliary-enhanced clause structure

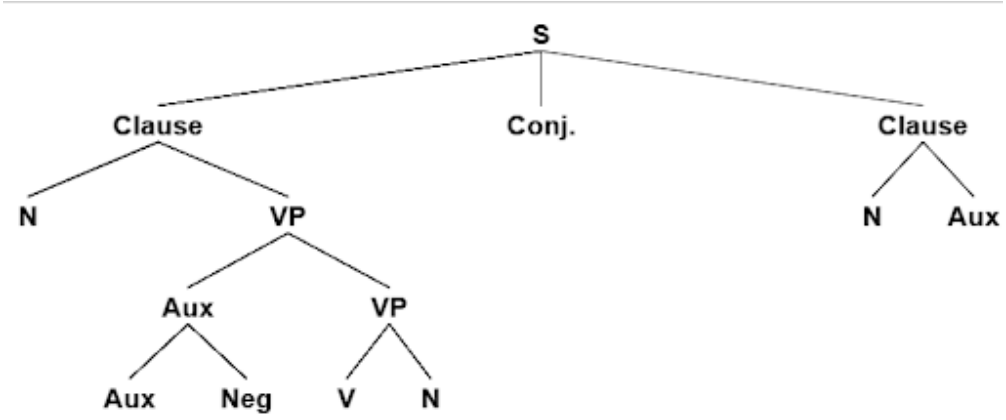


3. **Negation-Enhanced Clause Structure (type-2):**

The first clause contains both an auxiliary verb (*Aux*) and a negation element (*Neg*).The second clause remains elliptical. Example: *"Sally will not teach history, but Bill will."*

Figure 2.4

The syntactic tree of negation-enhanced clause structure



A key constraint in dataset construction is ensuring that contrast sentences maintain identical surface structures to elliptical sentences while differing in meaning. Therefore, sentences like *"Sally teaches history, but Bill does not hike."* are not included, as the verb "hike" introduces an additional syntactic variation. This constraint results in a dataset of **3,616** sentence triplets.

All sentences are generated by the Context-Free Grammar (CFG) generator from the Natural Language Toolkit (NLTK) in Python (Qi et al., 2020). A CFG is a type of formal grammar that defines the syntax of a language through a set of production rules. Each rule specifies how a symbol can be expanded into other symbols or terminal words, enabling the systematic generation of structured sentences. This method ensures that every sentence adheres to predefined grammatical patterns, making it an ideal tool for analyzing linguistic structure. NLTK, a widely used Python library for natural language processing, provides tools for generating and manipulating text based on CFGs. By implementing NLTK, this experiment maintains precise control over the surface structure of the sentences.

Additionally, the contrast sentences are divided into two groups based on lexical similarity to verbs in elliptical sentences:

- **Selected Verbs Group:** Contrast sentences use verbs semantically similar to those in elliptical sentences.
- **Random Verbs Group:** Contrast sentences use verbs with greater semantic distance from those in elliptical sentences.

Table 1

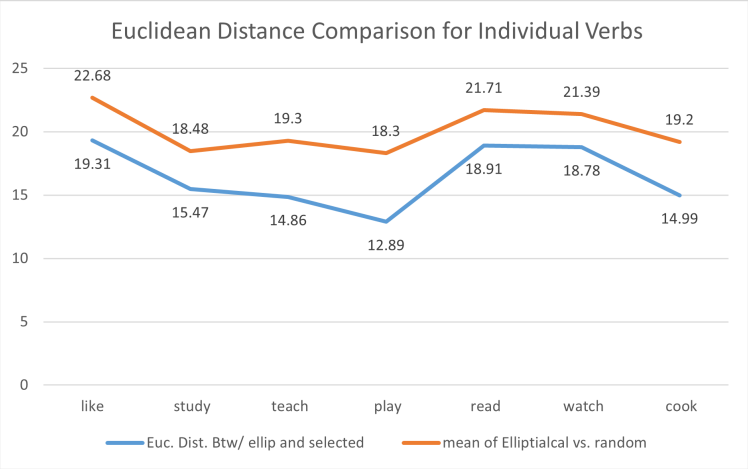
Table of verbs used in elliptical sentences, selected contrast, and random contrast

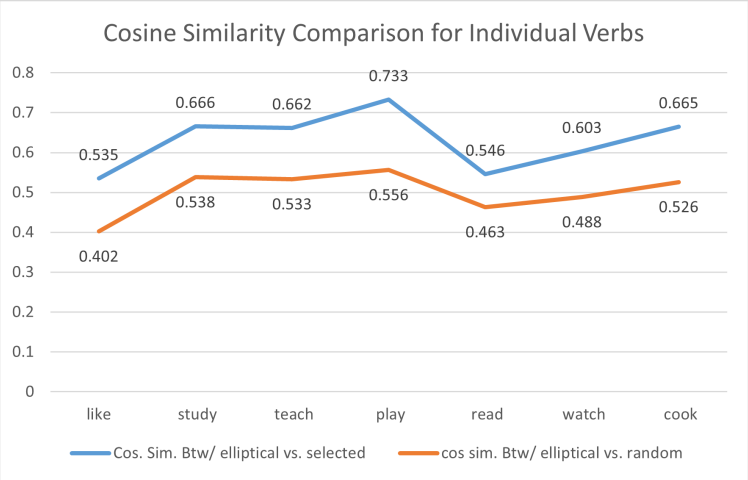
Verbs in elliptical/full (transitive)	Verbs in contrast (selected)	Verbs in contrast (random)
like	appreciate	swim
study	learn	run
teach	tutor	dance
play (chess, etc.)	participate	clap
read	interpret	snooze
watch	see	dream
cook	make	hike

To validate this distinction, the experiment measures Euclidean distance and cosine similarity between verb embeddings across groups. As expected, verbs in the selected group exhibit higher similarity to elliptical sentence verbs than those in the random group (see Figure 2.5). By controlling these variables, the dataset ensures that observed differences in BERT's embeddings result from compositionality rather than lexical co-occurrence patterns.

Figure 2.5

Graph of euclidean distance and cosine similarity between verbs in contrast sentences and verbs in elliptical sentences. The contrast verbs in the selected group are controlled to be more similar with those in elliptical than those in random group.





Measurement

The central hypothesis is that if BERT represents elliptical and full sentences as more similar than elliptical and contrast sentences, this would suggest that the model captures the underlying structure shared between elliptical and full sentences. Sentence pairs are passed through BERT to record their embedding vectors. With those vectors, the experiment evaluates sentence similarity by measuring three metrics:

Cosine Similarity: Measures semantic similarity by calculating the cosine of the angle between two vectors. Higher semantic similarity between two sentences leads to higher cosine simlairty between them. Thus, the cosine similarity between Full sentences and Elliptical sentences is expected to be high due to their semantic equivalence.

Cosine Similarity Ratio (X/Y): In addition to raw similarity scores, we compute a ratio that compares the similarity of elliptical–full pairs to the similarity of elliptical–contrast pairs. Specifically, we take

$$\text{Ratio} = \frac{X}{Y} = \frac{\text{CosineSim}(\text{Elliptical}, \text{Full})}{\text{CosineSim}(\text{Elliptical}, \text{Contrast})}.$$

A ratio greater than 1 indicates that elliptical sentences are closer to their full-sentence counterparts than to unrelated contrast sentences, suggesting that BERT captures the implicit verb phrase rather than relying solely on superficial lexical cues.

Euclidean Distance: To measure the overall difference between sentence embeddings, we calculate the Euclidean distance in the embedding space. While cosine similarity focuses on the angle between vectors, Euclidean distance considers both direction and magnitude, providing a holistic measure of how far apart the embeddings are in the high-dimensional space. This metric captures overall differences in embedding space, including both semantic and surface influences.

Expected Outcomes and Interpretation

It is expected that if BERT encodes ellipsis in a compositional way, elliptical sentences will be embedded similarly to their full-sentence counterparts—and distinctly from contrast sentences. In practice, this means that sentences like (1) should be closer in the embedding space to (2) than to a contrast sentence. If the ratio of elliptical–full cosine similarity to elliptical–contrast similarity (x/y) consistently exceeds 1, it would suggest that BERT detects the silenced structure. Additionally, it is also expected that **random contrast verbs**—which are semantically more distant from the elliptical verbs—will cause elliptical–contrast pairs to appear less similar, with **a higher elliptical–full vs. elliptical–contrast ratio (x/y)** than in the selected group. This would further support the idea that BERT encodes the implicit verb phrase, rather than being driven by superficial lexical overlap. Conversely, if the cosine similarity ratio (x/y) fails to exceed 1, it implies that BERT does not reliably detect the hidden structure in ellipsis.

Results

The results are shown in the foollowing tables. The first result shows that the cosine similarity ratio between elliptical–full pairs and elliptical–contrast pairs exceeds 1 (with the mean of 1.037), suggesting that BERT's embeddings of elliptical sentences tend to align more closely with their full counterparts. Notably, this ratio is higher for sentences in the random group (with the mean of 1.037) than those in the selected group (with the mean of 1.001), which aligns with our expectation. Euclidean distances for elliptical–full pairs, however, are often farther apart than expected (see Figure 3.1). Furthermore, no single word is found to significantly affect the observed outcomes, implying that the results are not driven by surface-level lexical choices (see

figures in Appendix 1).

Table 2
Results from the selected contrast group

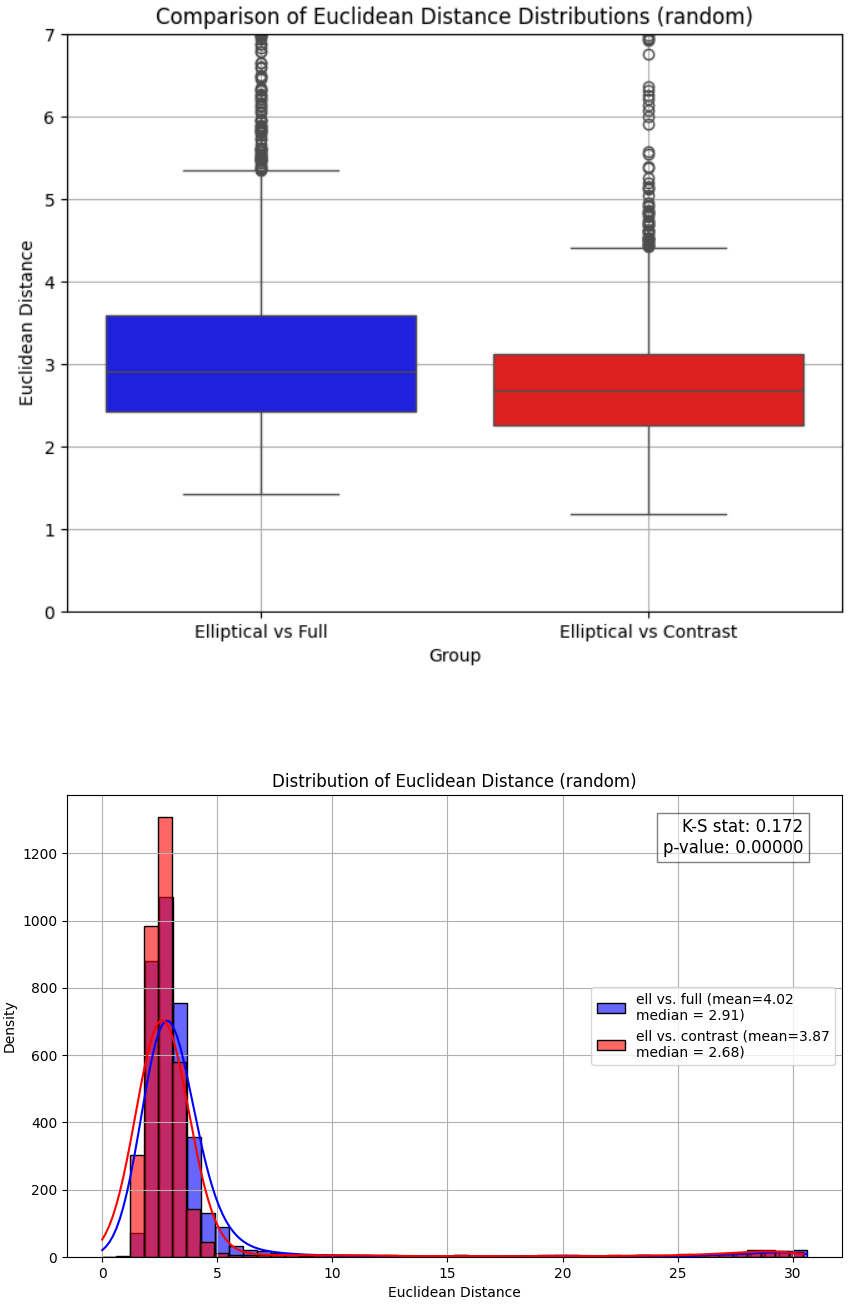
	Cos.Sim. (Ell_Full)	**Cos.Sim. (Ell_Con)**	**Cos. Sim. Ratio (X/Y)**
General	0.956	0.961	**1.001**
Basic Clause (type-0)	0.959	0.963	1.056
Auxiliary-Enhanced Clause (type-1)	0.924	0.936	0.987
Negation-Enhanced Clause (type-2)	0.986	0.987	0.999

Table 3
Results from the random contrast group

	Cos.Sim. (Ell_Full)	**Cos.Sim. (Ell_Con)**	**Cos. Sim. Ratio (X/Y)**
General	0.956	0.954	**1.037**
Basic Clause (type-0)	0.959	0.956	1.043
Auxiliary-Enhanced Clause (type-1)	0.924	0.922	1.064
Negation-Enhanced Clause (type-2)	0.986	0.986	1.006

Figure 3.1

Visualization of distribution of euclidean distance



Note. These two graphs show the distribution of Euclidean distances for each sentence pair type. Elliptical sentences are embedded further away from their full counterparts than contrast sentences.

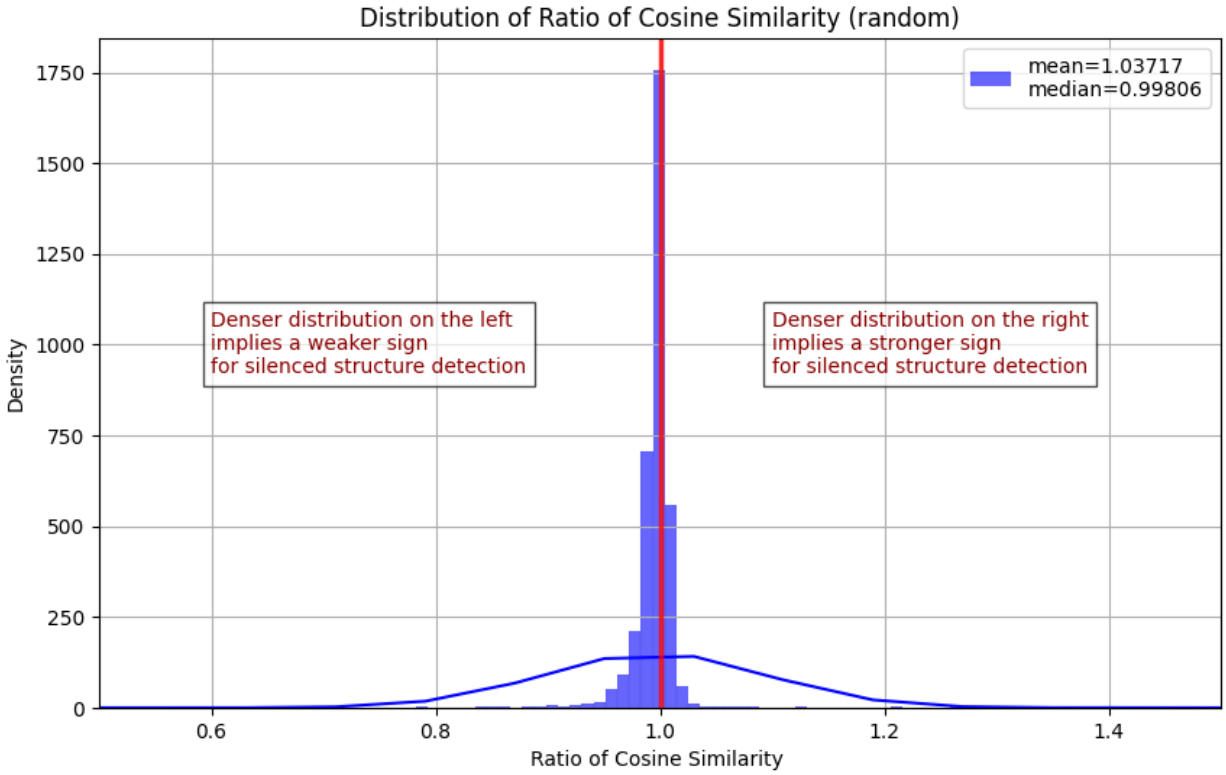
Consider the measurement of the cosine similarity ratio (X/Y):

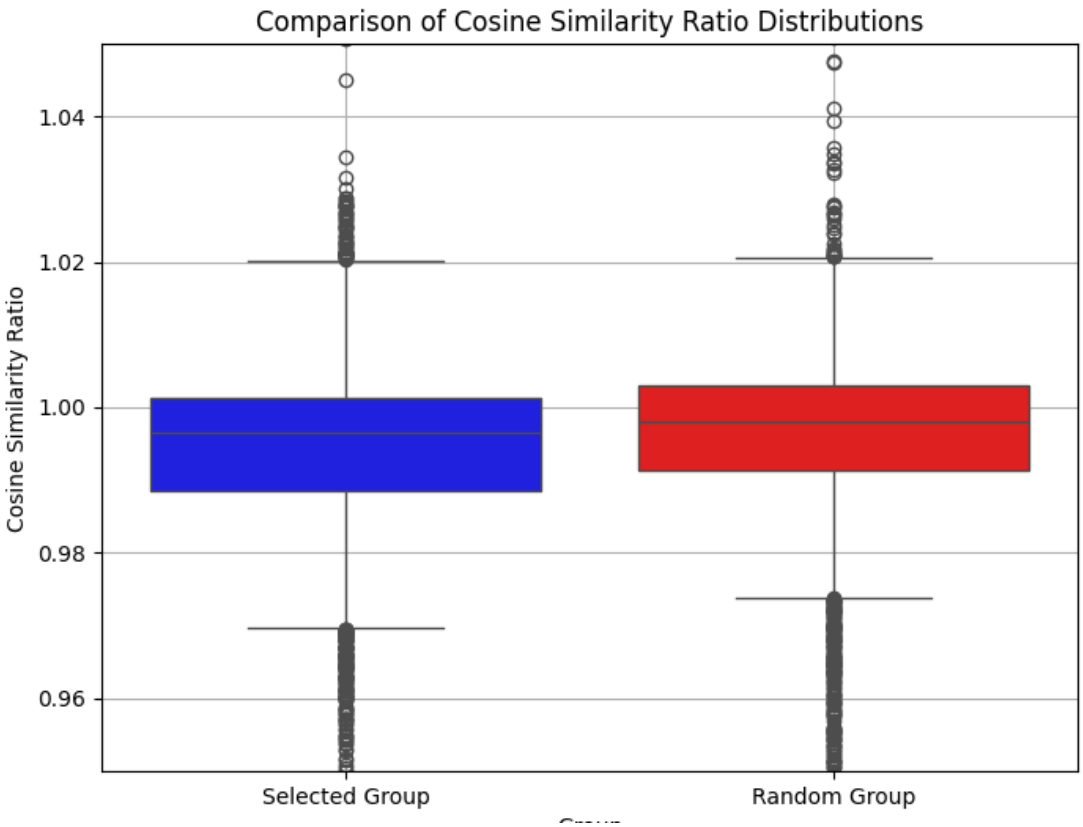
$$\text{Ratio} = \frac{X}{Y} = \frac{\text{CosineSim}(\text{Elliptical}, \text{Full})}{\text{CosineSim}(\text{Elliptical}, \text{Contrast})}.$$

If BERT fully captures the omitted elements, we would expect this ratio to be consistently **greater than 1**, indicating that elliptical–full pairs outscore elliptical–contrast pairs by a clear margin. The results, however, are mixed. On one hand, the mean ratio is indeed above 1, suggesting that elliptical sentences align more closely with their full-sentence counterparts. On the other hand, the ratio exhibits a wide distribution: the random group shows a standard deviation of 0.5. Moreover, the negation-enhanced clause structure (type-2) complicates the picture: its mean ratio falls slightly below 1, whereas the median is above 1, suggesting a cluster of outliers pulling the average downward. Additionally, different **sentence structure types** appear to demonstrate varying levels of influence on the ratio (see Figure 3.3), indicating that sentence structure does affect how BERT treats elliptical sentences and their full counterparts. There are no significant correlations emerged between the presence of specific words and any of the embedding-space metrics examined (Figure 3.4). In other words, the cosine similarity ratio does not appear to be systematically driven by particular words.

Figure 3.2

Graphs of cosine similarity ratio

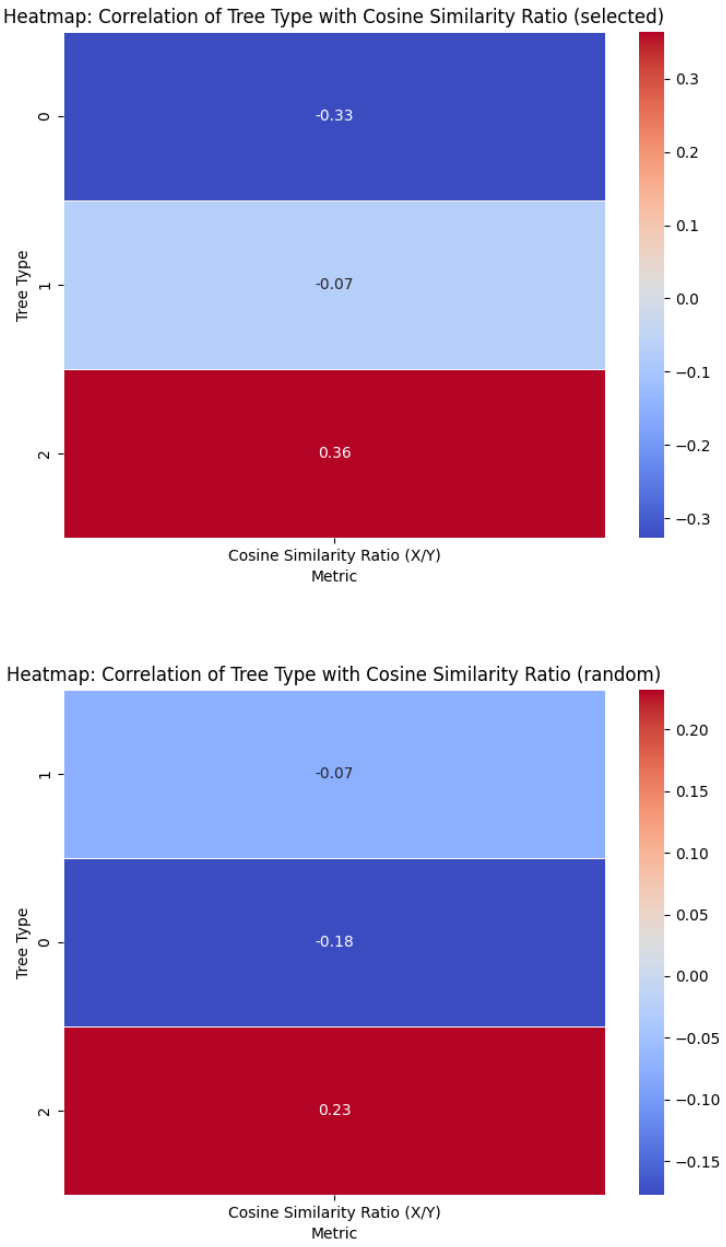


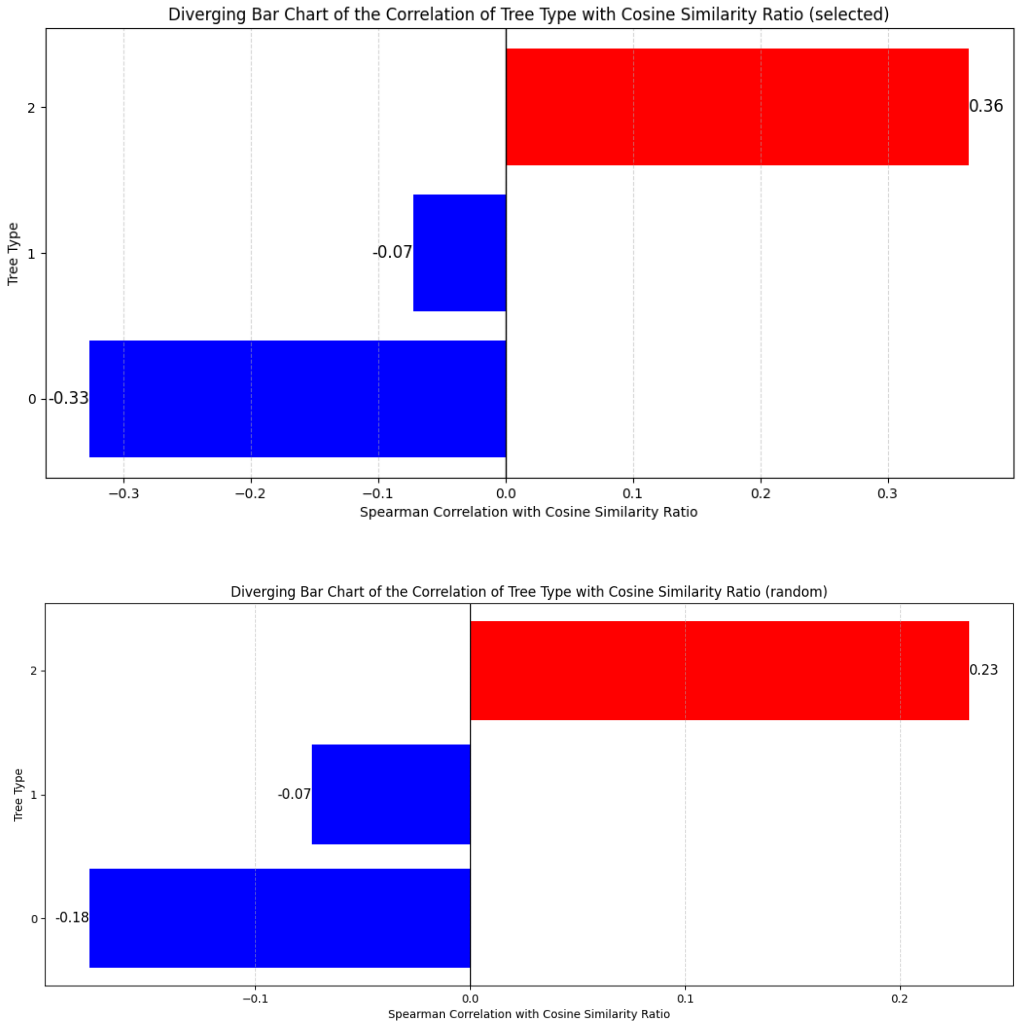


Note. The histogram shows the distribution of cosine similarity ratios, where values above 1.0 suggest sensitivity to underlying structure, and values below 1.0 suggest insensitivity. The boxplot compares selected and random contrast groups, indicating that BERT is sensitive to verb in sentences.

Figure 3.3

Heatmaps and Diverging Bar Plot of correlation between ratio and Structure Types

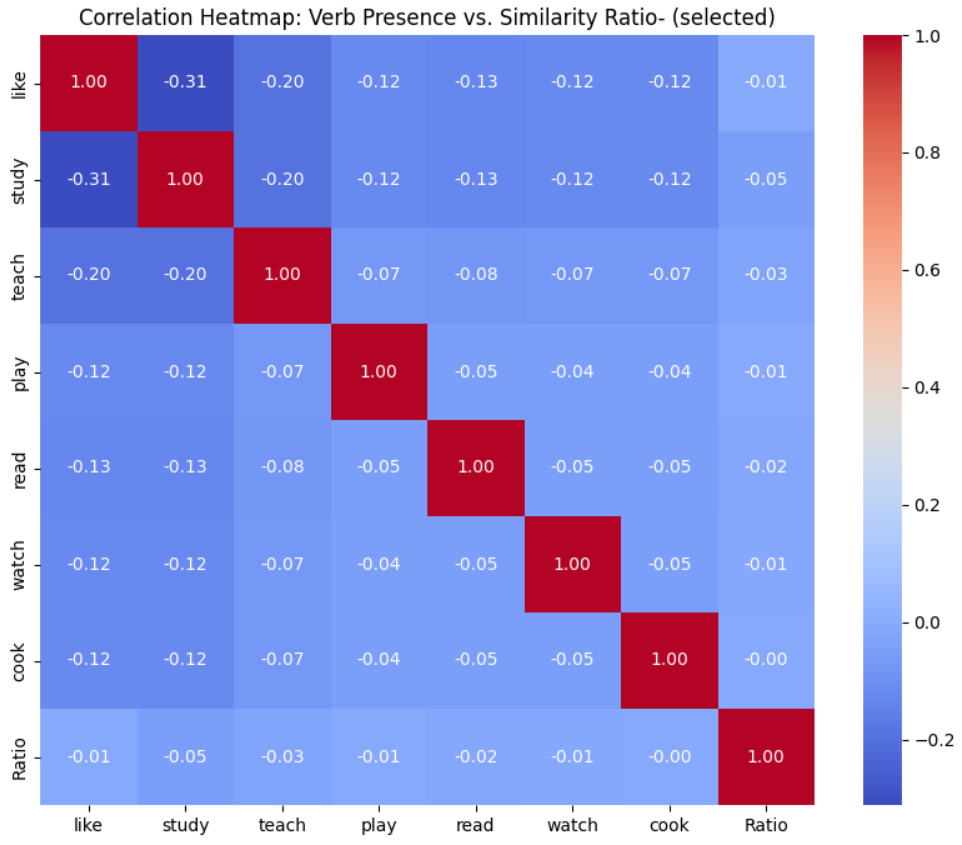


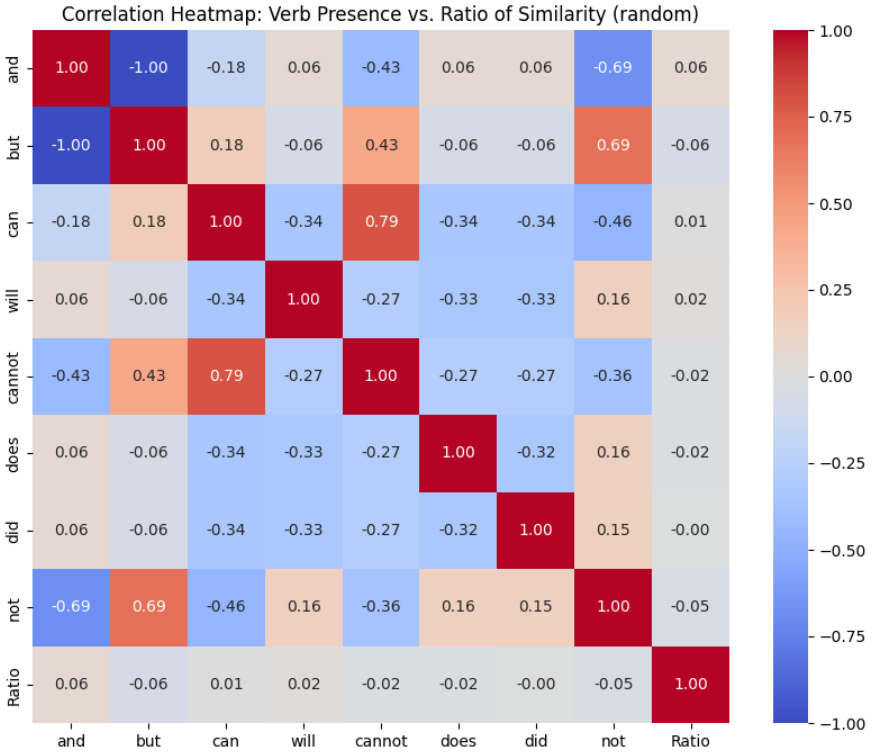


Note. These figures show the correlation between sentence structure types and the cosine similarity ratio. A value close to 1 indicates strong positive correlation, while a value close to -1 indicates strong negative correlation; values near 0 imply no correlation. The results show that different surface structures exhibit varying degrees of correlation with the ratio, and this pattern is observed across both the selected and random groups.

Figure 3.4

Heatmap of words and metrics





Note. These heatmaps show the correlation between individual verb presence and the cosine similarity ratio. A value close to 1 or -1 would indicate a strong correlation, while values near 0 suggest no correlation. As shown, no strong correlations are observed, suggesting that the similarity ratio is not driven by the presence of specific words. This supports the interpretation that the finding shows sensitivity to sentence structure rather than lexical use at surface.

Discussion

The results indicate that BERT detects silenced structures in elliptical sentences but not as strong or consistent as expected. Specifically, the cosine similarity ratio between elliptical–full pairs and elliptical–contrast pairs is higher than 1, with a mean value of **1.037**. This means that, on average, BERT's embeddings of elliptical sentences are more similar to their full counterparts than contrast sentences. However, the magnitude of this effect is small, and other variables complicate the case.

Signs for Compositional Processing

Mean Cosine Similarity Ratio greater than 1: The fact that the mean similarity ratio is greater than 1 suggests that BERT, to a certain extent, detects the silenced verb phrase in elliptical sentences rather than just considering them as independent fragments. This aligns with our hypothesis that BERT captures syntactic structures rather than relying solely on surface-level lexical patterns.

Stronger Effect in the Random Contrast Group: The higher ratio for the **random contrast group** (1.037) compared to the **selected contrast group** (1.001) shows that BERT is sensitive to the choice of contrast verbs. When contrast verbs are semantically unrelated to those in the elliptical sentence, BERT more clearly distinguishes elliptical sentences from contrast sentences. This result is in accordance with human language use—verbs that are more distant create a greater distinction in sentence. And this fact reinforces our hypothesis about BERT's compositionality.

No Single Word Drives the Effect: The absence of strong correlations between individual words and embedding-space metrics further suggests that BERT is not simply relying on surface-level lexical uses. If specific words had an outstanding impact on cosine similarity ratio, we might suspect that the model's behavior was due to statistical biases in the dataset rather than silenced structure detection. The lack of such an effect supports the BERT's compositionality.

Limitations and Unexpected Findings

Low Magnitude of the Effect: Although the mean ratio is above 1, the actual **numerical difference is small** (1.037 for the random group, 1.001 for the selected group). Ideally, if BERT were fully encoding the implicit elements of ellipsis, we would expect a more substantial separation between elliptical–full and elliptical–contrast similarity scores. The small effect size suggests that while BERT captures some compositionality, it does not do so with high consistency.

Effects by Sentence Structure: Unlike other clause structures, the **negation-enhanced clause structure (Type-2)** produces a mean ratio slightly below 1, though the median remains above 1. Moreover, the variation in correlation between different clause structures and the Cosine

Similarity Ratio (X/Y) indicates the clause structure impacts the way for machines to process sentences (see Figure 3.3). One possible explanation is that different clause structures inherently affect how verbs are represented.

A comparison of sample sentences across clause structures (see Table 4) reveals that sentences of the **Basic Clause Structure (Type-0) do not preserve the original verb form (e.g., "teach" vs. "teaches"), whereas the other two structures maintain verb consistency.** This structural difference may partly account for the observed effects of clause structure divergence. However, the precise extent and validity of this effect remain open questions for further investigation.

Table 4
Clause Structure Comparisons

Clause Structure Type	Sample Sentences
Basic Clause Structure (Type-0)	Sally teaches history, but John does not.
Auxiliary-Enhanced Clause Structure (Type-1)	Sally will teach history, and John will.
Negation-Enhanced Clause Structure (Type-2)	Sally will not teach history, but John will.

Dataset Size Constraints: The relatively small dataset (3,616 sentence triplets) might be a factor for the observed variability. A bigger dataset would likely give more stable and interpretable results, reducing the influence of outliers and improving statistical reliability.

Potential Limited Distinciton between Verbs in Elliptical and Random Contrast Sentences: One possible explanation for the **low magnitude of the ratio** is that the verbs in elliptical sentences and their random contrast counterparts are **not sufficiently distinct**. While the random group was designed to introduce contrast, many of them may still share some **s**imilarity with the original elliptical verbs. This overlap may have minimized the expected contrast effect. If contrast verbs were even more distinct in meaning, we might have observed a higher ratio score.

Conclusions

This paper investigates whether BERT encodes compositionality by analyzing its embedding of VP-ellipsis. And the results suggest that BERT demonstrates some sensitivity to compositional relations, as indicated by the mean cosine similarity ratio exceeding 1. This finding suggests that elliptical sentences are embedded more similarly to their full-sentence counterparts than to unrelated contrast sentences. However, the extent of this effect is small, and there is notable variability across different sentence structures. In particular, the **negation-enhanced clause structure (Type-2)** presents a specific challenge, with a mean ratio slightly below 1.

The observation that BERT encodes compositionality to some extent aligns with existing research showing that LLMs capture hierarchical syntactic structures (Bowman et al., 2015), which may assist LLMs' detection of silent structure in ellipsis. Moreover, the mean cosine similarity ratio exceeding 1 supports previous findings that LLMs can grasp essential aspects of semantic meaning (Liang et al., 2015; Pavlick et al., 2023). The sign of compositionality is particularly noteworthy given prior studies showing LLMs' failure in ellipsis-related natural language tasks (Hardt, 2023).

At the same time, the small effect size and the variation in BERT's encoding of VP-ellipsis across different sentence structures suggests that compositionality in LLMs is not fully robust. Prior research has highlighted that LLMs do not rely on discrete symbols in the same way human cognition does (Biever et al., 2023; Bishop, 2021; Lopes et al., 2023; Nefdt et al., 2022), which may contribute to the observed instability. Additionally, the lack of explicitly rule-based operations in LLMs may contribute to the observed inconsistencies in ellipsis processing across different sentence structures (Lake, 2018; Baron, 2019). Nonetheless, rather than negating compositionality in LLMs altogether, these factors suggest that compositionality in LLMs is a nuanced property that requires further detailed investigations.

Beyond its empirical findings, this study also introduces a novel methodological approach. Unlike prior research that evaluates LLMs through generated outputs, this study directly examines their internal vector representations. By focusing on embedding-space

measurements, this approach isolates structural encoding from confounding factors such as training data biases or text generation refinement. This distinction is crucial for determining whether compositionality in LLMs stems from genuine structural sensitivity or merely from effective output training.

Moving forward, several directions remain open for further research:

- **Expanding the Dataset**: A limitation of this study is the relatively small dataset of 3,616 sentence triplets. Increasing the dataset size could help produce more stable and predictable results.
- **Exploring Different Types of Hidden Structure beyond VP-Ellipsis**: This study focuses exclusively on VP-Ellipsis, but ellipsis occurs in various forms which may behave differently within BERT's embeddings. Further research could investigate how BERT encodes the following structures:
  - **Sluicing:** Cases where the entire clause is omitted except for the question word. For example, in *"Someone left early, but I don't know who,"* the clause *"left early"* is omitted after the question word *"who."*
  - **Gapping:** Sentences where the second clause contrasts with elements in the first clause. For instance, in *"John can play the guitar, and Mary the violin,"* the verb *"play"* is omitted in the second clause but is understood from the first.
  - **Anaphora:** Sentences where a word or phrase refers back to something previously mentioned, requiring contextual reconstruction. For example, in *"Sarah loves painting. She does it everyday,"* the pronoun *"it"* refers back to *"painting"*, which is not explicitly stated.
- **Different Verbs and Sentences**: Although no single word was found to significantly affect the similarity ratio, the correlation between specific verbs and compositional effects varied. Future studies could explore a broader range of **modal verbs** (*could, would, should, shall*),

which might introduce different compositional hierarchy within sentences. It is also worthy to consider **stronger semantic contrasts** in contrast sentences so that verbs in the random contrast group are sufficiently distinct from their elliptical counterparts. In addition, given that sentence structure has a certain impact on LLMs' interpretation of sentences, future studies should explore a more diverse range of sentence structures to assess the generalizability of these findings.

- **Revisiting the Structural Approach to Ellipsis**: This paper assumes the structural approach to ellipsis, in which the omitted part is present in the syntactic structure but silenced. The findings, however, suggest that BERT may not fully reconstruct these silenced structures. This raises the question of whether **semantic or discourse-based approaches**—which emphasize contexts—might better account for BERT's performance. Future experiments, with new set-ups, could explore LLMs' performance based on different linguistic theories.

Overall, this paper provides evidence that BERT encodes compositionality to an unstable extent. Whether this reflects a limitation in transformer-based LLMs or fundamental constraints in compositionality is unclear. Only with more detailed investigations would we determine to what extent LLMs capture the compositional nature of human language and thought.

References

3Blue1Brown. (2017, November 3). What is backpropagation really doing? | Chapter 3, deep learning [Video]. YouTube. https://www.youtube.com/watch?v=Ilg3gGewQ5U

Block, N. (1996). Mental paint and mental latex. *Philosophical Issues, 7*, 19–49. https://doi.org/10.2307/1522889

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, *2*(3), 113–124. https://doi.org/10.1109/tit.1956.1056813

Chomsky, N. (2002). *Syntactic structures* (2nd ed., D. W. Lightfoot, Intro.). Berlin & New York: Mouton de Gruyter.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (pp. 493–522). Wiley Blackwell. https://doi.org/10.1002/9781118882139.ch16

Clark, A. (2017). Computational learning of syntax. *Annual Review of Linguistics, 3*(1), 107–123. https://doi.org/10.1146/annurev-linguistics-011516-034008

Davidson, D. (1967). Truth and Meaning. *Synthese (Dordrecht), 17*(3), 304323. https://doi.org/10.1007/BF00485035

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805

Dretske, F. (1993). Conscious experience. *Mind, 102*(406), 263–283. https://doi.org/10.1093/mind/102.406.263

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass, 6*(10), 635–653. https://doi.org/10.1002/lnco.362

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, MA: The MIT Press.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited* (1st ed.). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199548774.001.0001

Gennaro, R. J. (2008). Representationalism, peripheral awareness, and the transparency of experience. *Philosophical Studies, 139*(1), 39–56. https://doi.org/10.1007/s11098-007-9101-4

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the transformer-based models for NLP tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)* (pp. 179–183). Polish Information Processing Society. https://doi.org/10.15439/2020F20

Grossberg, S. (2013). Recurrent neural networks. *Scholarpedia Journal, 8*(2), 1888. https://doi.org/10.4249/scholarpedia.1888

Hardt, D. (2023). Ellipsis-dependent Reasoning: A New Challenge for Large Language Models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 39-47). Association for Computational Linguistics. https://aclanthology.org/volumes/2023.acl-short/

Heinz, J., Snyder, W., Pater, J., & Lidz, J. (2016). Computational theories of learning and developmental psycholinguistics. In J. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford handbook of developmental linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199601264.013.27

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural network generalise? *Journal of Artificial Intelligence Research, 67*, 757–795. https://doi.org/10.1613/jair.1.11674

Kazanina, N., & Poeppel, D. (2023). The neural ingredients for a language of thought are available. *Trends in Cognitive Sciences, 27*(11), 996–1007. https://doi.org/10.1016/j.tics.2023.07.012

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Levin, J. (2023, April 4). Functionalism. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/functionalism/

Lew, A. K., Tessler, M. H., Mansinghka, V. K., & Tenenbaum, J. B. (2020). Leveraging unstructured statistical knowledge in a probabilistic language of thought. *Proceedings of the Annual Meeting of the Cognitive Science Society, 42*.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-Centric Learning with Slot Attention. *arXiv.Org*.

Lurz, R. (2012). Uriah Kriegel, *Subjective consciousness: A self-representational theory* (Oxford University Press, 2009). In *Bolzano & Kant, 85*, 347–353. https://doi.org/10.1163/9789401208338_020

Lycan, W. (2023, October 19). Representational theories of consciousness. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/consciousness-representational/

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Merchant, J., Temmerman, T., & van Craenenbroeck, J. (2018). Ellipsis: A survey of analytical approaches. In *The Oxford Handbook of Ellipsis*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198712398.013.2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. https://arxiv.org/abs/1301.3781

Pagin, P. (2003). Communication and strong compositionality. *Journal of Philosophical Logic*, *32*(3), 287–322. https://doi.org/10.1023/a:1024258529030

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95*(1), e41–e74. https://doi.org/10.1353/lan.2019.0009

Partee, B. H. (2001). Montague grammar. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 9995–9999). https://doi.org/10.1016/b0-08-043076-7/02954-5

Petersen, E., & Potts, C. (2023). Lexical semantics with large language models: A case study of English "break." *Findings of the Association for Computational Linguistics: EACL 2023*, 490–511. https://doi.org/10.18653/v1/2023.findings-eacl.36

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences, 46*, e261. https://doi.org/10.1017/S0140525X22002849

Rescorla, M. (2020, February 21). The computational theory of mind. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/computational-mind/

Rescorla, M. (2023, October 16). The language of thought hypothesis. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/language-thought/

Rio-Torto, I., Fernandes, K., & Teixeira, L. F. (2020). Understanding the decisions of CNNs: An in-model approach. *Pattern Recognition Letters, 133*, 373–380. https://doi.org/10.1016/j.patrec.2020.04.004

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386-408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 216-271). MIT Press.

Tye, M. (1997). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.

Tye, M. (2003). A theory of phenomenal concepts. *Royal Institute of Philosophy Supplement, 53*,

91–105. https://doi.org/10.1017/S1358246100008286

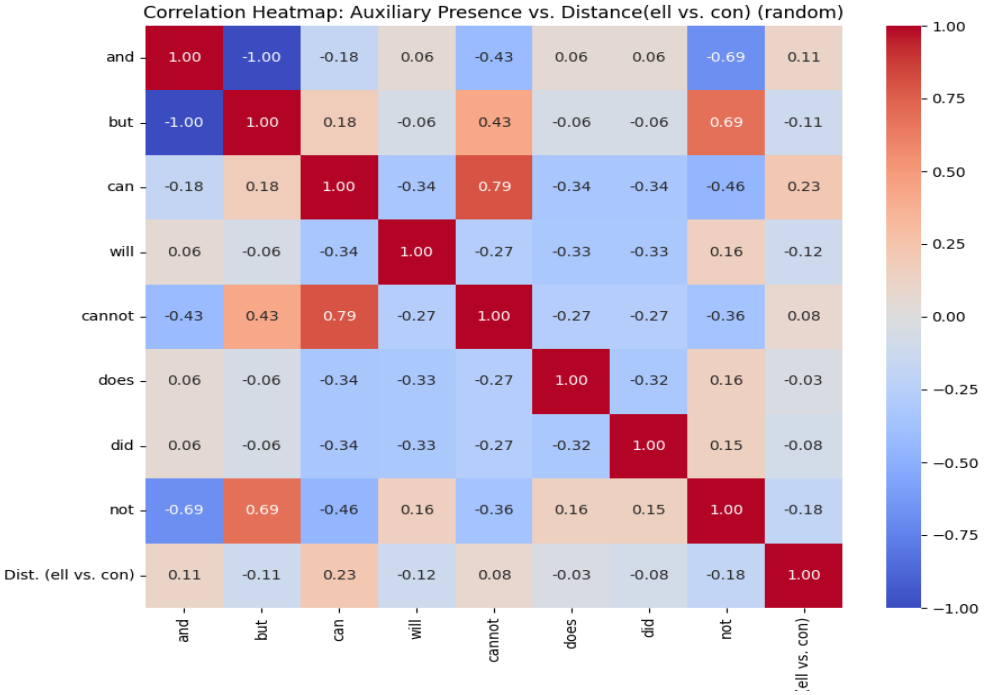
Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020b, April 23). *Stanza: A python natural language processing toolkit for many human languages*. arXiv.org. https://arxiv.org/abs/2003.07082

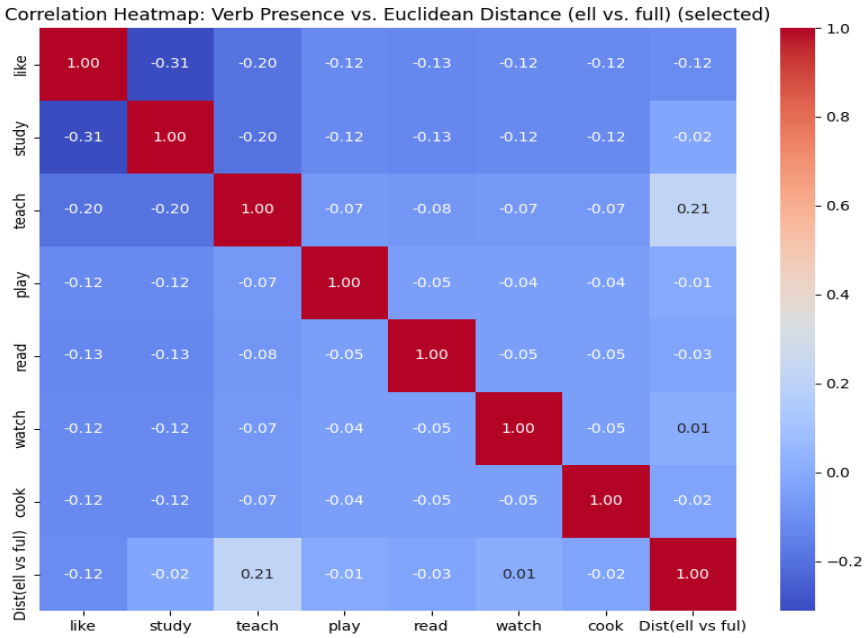
Zou, J., Han, Y., & So, S.-S. (2008). Overview of artificial neural networks. In *Artificial Neural Networks, 458*, 14–22. https://doi.org/10.1007/978-1-603

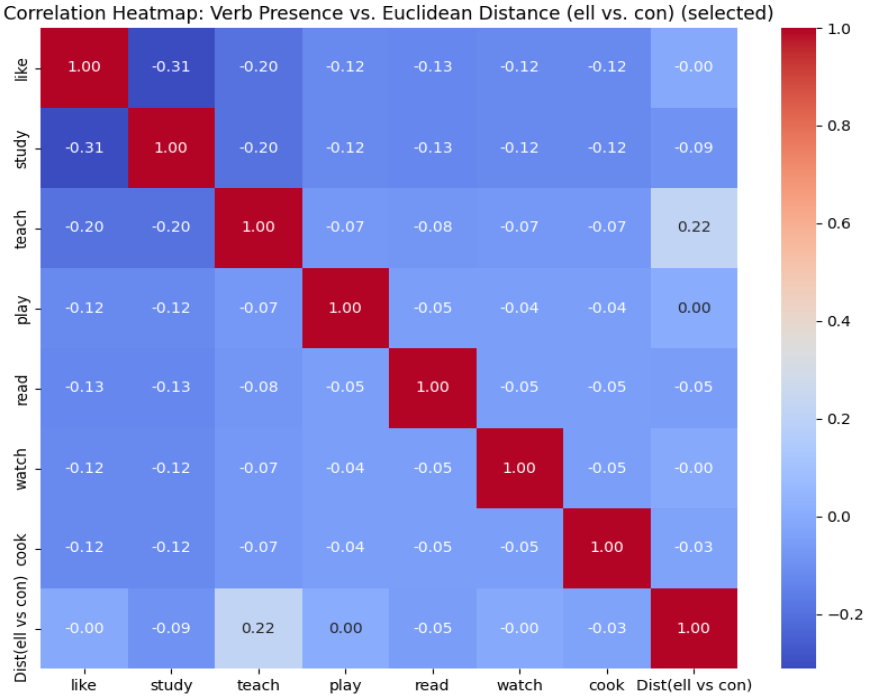
Appendix

Figure 1

Heatmap of correlation between words and euclidean distance between sentences






Correlation Heatmap: Verb Presence vs. Euclidean Distance (ell vs. con) (selected)
like
study
teach
play
read
watch
cook
Dist(ell vs con)
1.00 -0.31 -0.20 -0.12 -0.13 -0.12 -0.12 -0.00
-0.31 1.00 -0.20 -0.12 -0.13 -0.12 -0.12 -0.09
-0.20 -0.20 1.00 -0.07 -0.08 -0.07 -0.07 0.22
-0.12 -0.12 -0.07 1.00 -0.05 -0.04 -0.04 0.00
-0.13 -0.13 -0.08 -0.05 1.00 -0.05 -0.05 -0.05
-0.12 -0.12 -0.07 -0.04 -0.05 1.00 -0.05 -0.00
-0.12 -0.12 -0.07 -0.04 -0.05 -0.05 1.00 -0.03
-0.00 -0.09 0.22 0.00 -0.05 -0.00 -0.03 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

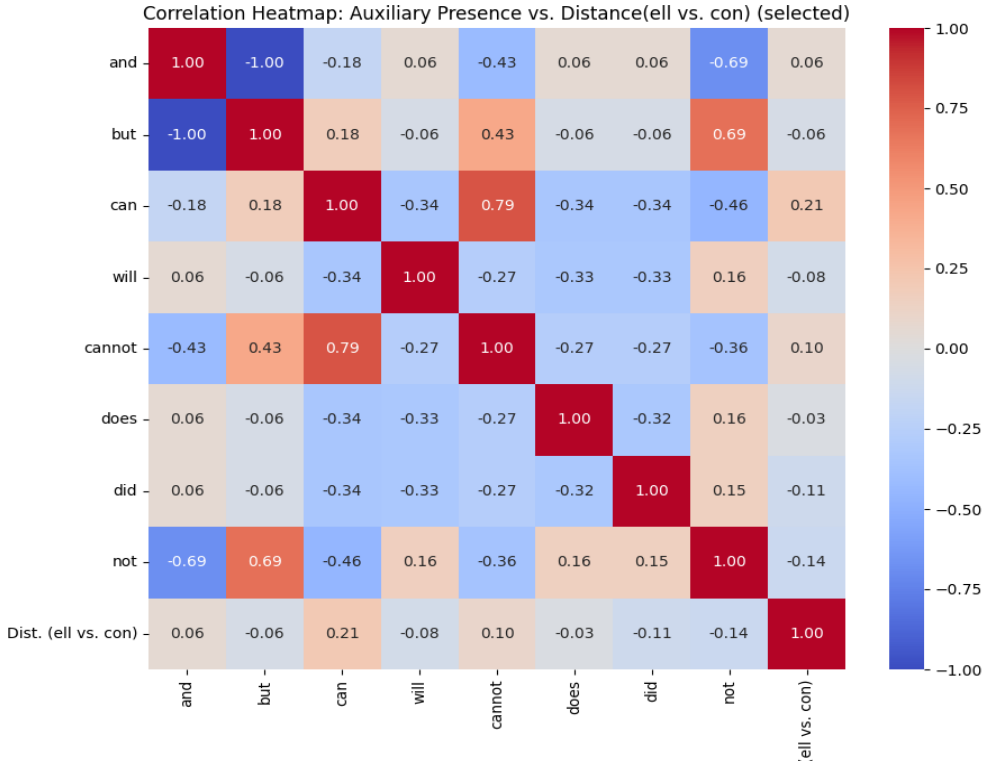
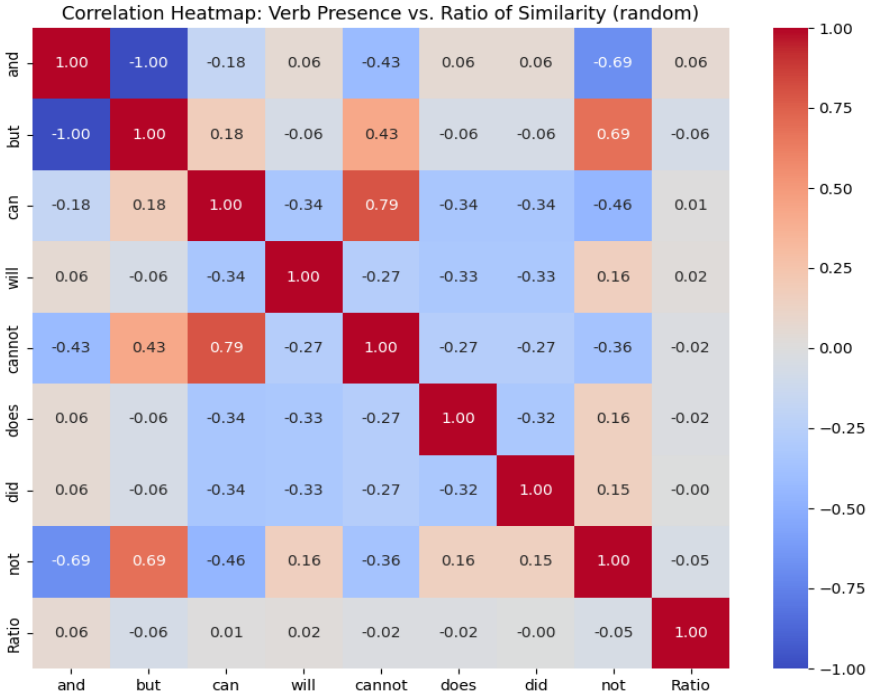

Correlation Heatmap: Auxiliary Presence vs. Distance(ell vs. con) (selected)
and
but
can
will
cannot
does
did
not
Dist. (ell vs. con)
1.00 -1.00 -0.18 0.06 -0.43 0.06 0.06 -0.69 0.06
-1.00 1.00 0.18 -0.06 0.43 -0.06 -0.06 0.69 -0.06
-0.18 0.18 1.00 -0.34 0.79 -0.34 -0.34 -0.46 0.21
0.06 -0.06 -0.34 1.00 -0.27 -0.33 -0.33 0.16 -0.08
-0.43 0.43 0.79 -0.27 1.00 -0.27 -0.27 -0.36 0.10
0.06 -0.06 -0.34 -0.33 -0.27 1.00 -0.32 0.16 -0.03
0.06 -0.06 -0.34 -0.33 -0.27 -0.32 1.00 0.15 -0.11
-0.69 0.69 -0.46 0.16 -0.36 0.16 0.15 1.00 -0.14
0.06 -0.06 0.21 -0.08 0.10 -0.03 -0.11 -0.14 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 2

Heatmap for the correlation between words and cosine similarity ratio (x/y)


Correlation Heatmap: Verb Presence vs. Ratio of Similarity (random)
and 1.00 -1.00 -0.18 0.06 -0.43 0.06 0.06 -0.69 0.06
but -1.00 1.00 0.18 -0.06 0.43 -0.06 -0.06 0.69 -0.06
can -0.18 0.18 1.00 -0.34 0.79 -0.34 -0.34 -0.46 0.01
will 0.06 -0.06 -0.34 1.00 -0.27 -0.33 -0.33 0.16 0.02
cannot -0.43 0.43 0.79 -0.27 1.00 -0.27 -0.27 -0.36 -0.02
does 0.06 -0.06 -0.34 -0.33 -0.27 1.00 -0.32 0.16 -0.02
did 0.06 -0.06 -0.34 -0.33 -0.27 -0.32 1.00 0.15 -0.00
not -0.69 0.69 -0.46 0.16 -0.36 0.16 0.15 1.00 -0.05
Ratio 0.06 -0.06 0.01 0.02 -0.02 -0.02 -0.00 -0.05 1.00
and but can will cannot does did not Ratio
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

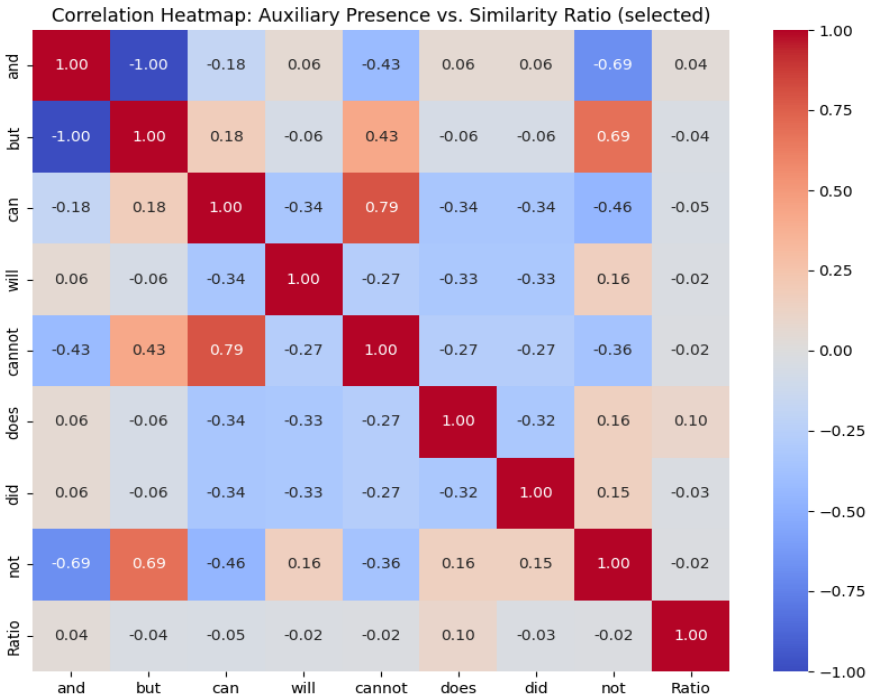

Correlation Heatmap: Auxiliary Presence vs. Similarity Ratio (selected)
and 1.00 -1.00 -0.18 0.06 -0.43 0.06 0.06 -0.69 0.04
but -1.00 1.00 0.18 -0.06 0.43 -0.06 -0.06 0.69 -0.04
can -0.18 0.18 1.00 -0.34 0.79 -0.34 -0.34 -0.46 -0.05
will 0.06 -0.06 -0.34 1.00 -0.27 -0.33 -0.33 0.16 -0.02
cannot -0.43 0.43 0.79 -0.27 1.00 -0.27 -0.27 -0.36 -0.02
does 0.06 -0.06 -0.34 -0.33 -0.27 1.00 -0.32 0.16 0.10
did 0.06 -0.06 -0.34 -0.33 -0.27 -0.32 1.00 0.15 -0.03
not -0.69 0.69 -0.46 0.16 -0.36 0.16 0.15 1.00 -0.02
Ratio 0.04 -0.04 -0.05 -0.02 -0.02 0.10 -0.03 -0.02 1.00
and but can will cannot does did not Ratio
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00

Table 5
Summary Statistics for Selected Contrast Group

Tree Type	Metric	Mean	Median	Std Dev	Min	Max
Overall	Cosine Sim (Ell vs Full)	0.955602	0.98543	0.135328	0.047948	0.996535
Overall	Cosine Sim (Ell vs Contrast)	0.961409	0.990254	0.133016	0.044413	0.997998
Overall	Cosine Sim Ratio (X/Y)	1.008805	0.996457	0.307759	0.089348	8.359788
Overall	Euclidean Dist (Ell vs Full)	4.016922	2.912273	4.63255	1.421071	30.61029
Overall	Euclidean Dist (Ell vs Contrast)	3.519269	2.373122	4.70539	1.069287	30.51445
0	Cosine Sim (Ell vs Full)	0.959388	0.981064	0.114285	0.086365	0.99518
0	Cosine Sim (Ell vs Contrast)	0.962918	0.992591	0.131233	0.117777	0.997998
0	Cosine Sim Ratio (X/Y)	1.056467	0.990397	0.602422	0.089348	8.359788
0	Euclidean Dist (Ell vs Full)	4.065484	3.297953	3.985593	1.649626	29.99409
0	Euclidean Dist (Ell vs Contrast)	3.326627	2.064333	4.856412	1.069287	29.65412
1	Cosine Sim (Ell vs Full)	0.92381	0.984269	0.192848	0.047948	0.996033
1	Cosine Sim (Ell vs Contrast)	0.93612	0.990323	0.183138	0.044413	0.996868
1	Cosine Sim Ratio (X/Y)	0.987383	0.995865	0.116524	0.136945	3.849126
1	Euclidean Dist (Ell vs Full)	5.199492	3.405075	6.487484	1.506389	30.61029
1	Euclidean Dist (Ell vs Contrast)	4.446789	2.376063	6.344213	1.310556	30.51445
2	Cosine Sim (Ell vs Full)	0.986319	0.988459	0.008679	0.867561	0.996535
2	Cosine Sim (Ell vs Contrast)	0.986836	0.98892	0.012446	0.61061	0.995905
2	Cosine Sim Ratio (X/Y)	0.999687	1.000036	0.019629	0.921689	1.623797
2	Euclidean Dist (Ell vs Full)	2.749147	2.598803	0.741583	1.421071	10.22763
2	Euclidean Dist (Ell vs Contrast)	2.677416	2.52189	0.815007	1.530715	19.77198

Table 6
Summary Statistics for Random Contrast Group

Tree Type	Metric	Mean	Median	Std Dev	Min	Max
Overall	Cosine Sim (Ell vs Full)	0.955602	0.98543	0.135328	0.047948	0.996535
Overall	Cosine Sim (Ell vs Contrast)	0.954191	0.987587	0.147491	0.050397	0.997568
Overall	Cosine Sim Ratio (X/Y)	1.037166	0.998059	0.502881	0.072808	15.95057
Overall	Euclidean Dist (Ell vs Full)	4.016922	2.912273	4.63255	1.421071	30.61029
Overall	Euclidean Dist (Ell vs Contrast)	3.870705	2.68114	5.107028	1.178978	30.2464
0	Cosine Sim (Ell vs Full)	0.959388	0.981064	0.114285	0.086365	0.99518
0	Cosine Sim (Ell vs Contrast)	0.956167	0.987701	0.138991	0.097322	0.99722
0	Cosine Sim Ratio (X/Y)	1.042707	0.994649	0.505263	0.318246	15.95057
0	Euclidean Dist (Ell vs Full)	4.065484	3.297953	3.985593	1.649626	29.99409
0	Euclidean Dist (Ell vs Contrast)	3.8351	2.669019	4.880072	1.366071	29.90412
1	Cosine Sim (Ell vs Full)	0.92381	0.984269	0.192848	0.047948	0.996033
1	Cosine Sim (Ell vs Contrast)	0.922034	0.987453	0.203521	0.050397	0.996624
1	Cosine Sim Ratio (X/Y)	1.063855	0.997489	0.664316	0.072808	15.95057
1	Euclidean Dist (Ell vs Full)	5.199492	3.045075	6.487484	1.506389	30.61029
1	Euclidean Dist (Ell vs Contrast)	5.046516	2.704432	6.969468	1.358557	30.4643
2	Cosine Sim (Ell vs Full)	0.986319	0.988459	0.008679	0.867561	0.996535
2	Cosine Sim (Ell vs Contrast)	0.986488	0.987663	0.026761	0.104528	0.997568
2	Cosine Sim Ratio (X/Y)	1.005617	1.000368	0.230272	0.917239	3.939713
2	Euclidean Dist (Ell vs Full)	2.749147	2.598803	0.741583	1.421071	10.22763
2	Euclidean Dist (Ell vs Contrast)	2.665619	2.669926	1.102649	1.178978	30.2464