문학석사 학위논문

Probing the Linguistic Knowledge of BERT based on the Layer-wise Investigation with Affinity Prober

Affinity Prober를 활용한 BERT 언어 지식의 레이어 별 탐침 연구

2023 년 08 월

서울대학교 대학원

언어학과 언어학전공

장 동 준

Probing the Linguistic Knowledge of BERT based on the Layer-wise Investigation with Affinity Prober

지도 교수 신 효 필

이 논문을 문학석사 학위논문으로 제출함
2023 년 08 월

서울대학교 대학원
언어학과 언어학전공
장 동 준

장동준의 문학석사 학위논문을 인준함
2023 년 08 월

위 원 장 ______ 이 상 아 ______ (인)

부위원장 ______ 신 효 필 ______ (인)

위 원 ______ 김 문 형 ______ (인)

Probing the Linguistic Knowledge of BERT based on the Layer-wise Investigation with Affinity Prober

Advising Professor, Dr. Hyopil Shin

Submitting a master's thesis of Art

August 2023

Graduate School of Humanities
Seoul National University
Linguistics Major

Dongjun Jang

Confirming the master's thesis written by
Dongjun Jang
August 2023

Chair ____Sangah Lee____(Seal)

Vice Chair ____Hyopil Shin____(Seal)

Examiner ____Munhyong Kim____(Seal)

Abstract

Jang, Dongjun

Department of Linguistics

The Graduate School

Seoul National University

This paper presents a comprehensive investigation into the linguistic knowledge embedded within BERT, a pre-trained language model based on the Transformer architecture. We reinforce and expand upon the methodology proposed by Jang et al (2022) by introducing the ADTRAS algorithm (An Algorithm for Decrypting Token Relationships within Attention Scores), which decrypts token relationships within BERT's attention scores to analyze patterns at each layer. Our experiments using ADTRAS algorithm demonstrate that BERT autonomously learns part-of-speech information by leveraging lexical categories. We also provide insights into the general tendencies of BERT's layers when processing content words and function words. Additionally, we introduce the Classification of Sentence Sequencing (CSS) as a Finetuning Strategy, enabling indirect learning from minimal pairs, and leverage the Affinity Prober to examine syntactic linguistic phenomena using the BLiMP dataset. By tracing patterns and clustering similar phenomena, we enhance our understanding of BERT's interpretation of linguistic structures. Furthermore, we establish in detail the attributes of BERT layers related to lexical categories by connecting the general tendencies of the layers generalized by the

ADTRAS algorithm with the results obtained through the Affinity Prober. Our study makes several contributions. First, we introduce the ADTRAS algorithm, which enables a comprehensive analysis of BERT's linguistic knowledge. Second, we provide experimental evidence demonstrating BERT's ability to learn part-of-speech information. Third, we offer insights into the tendencies observed in different layers of BERT. Fourth, we propose the CSS Finetuning Strategy, which allows for indirect learning from minimal pairs. Fifth, we successfully cluster syntactic phenomena using the Affinity Prober. Finally, we uncover the general attention tendency of BERT towards lexical categories.

Keyword : Natural Language Processing, BERT, linguistic knowledge, ADTRAS algorithm, part-of-speech, lexical categories, layer tendencies, content words, function words, Classification of Sentence Sequencing (CSS), Finetuning Strategy, Affinity Prober, syntactic linguistic phenomena, BLiMP dataset
Student Number : 2021-22754

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

In recent years, Natural Language Processing (NLP) has seen remarkable progress thanks to the introduction of deep learning-based pre-trained models. These models have captured considerable interest, largely due to the revolutionary Transformer architecture proposed by Vaswani et al. (2017). This groundbreaking architecture has opened doors for the creation of advanced models that leverage the power of the multi-layer Self Attention Mechanism. These models integrate various components, including the Multi-head Attention Layer.

One prominent example of such models is BERT, which was introduced by Devlin et al. (2019). BERT is a pre-trained language model based on the Transformer's encoder structure and has been trained using a cloze test-based method. This approach has positioned BERT as a specialized language model for Natural Language Understanding (NLU), outperforming existing neural network models on standard NLU benchmarks. BERT's performance is particularly noteworthy in challenging tasks like CoLA, where traditional neural network models face significant difficulties. The remarkable performance exhibited by BERT implies the existence of latent linguistic knowledge within BERT.

The field of BERTology (Rogers et al., 2020) has emerged through ongoing research, aiming to uncover the potential latent linguistic knowledge embedded within BERT. BERTology primarily focuses on investigating the depths of BERT's language processing capabilities and exploring the replication of language structures. Research in this area ranges from examining the model's post-training performance on language information (such as part-of-speech and named entities) to investigating the operational processes of language models, such as the self-attention mechanism, in order to reproduce syntactic structures or word dependencies. However, these approaches have limitations in terms of directly injecting language knowledge into the model to explore linguistic knowledge. The discussion of appropriately training the model with directly injected language information is an engineering topic. This means that it is not easy to investigate the inherent pure language knowledge within the language model. Therefore, in order to comprehensively investigate the linguistic knowledge embedded within BERT, it is essential to employ research methodologies that involve analyzing the model's outputs, such as embeddings and attention scores, generated during its computational process. These outputs should be interpreted from a linguistic perspective to uncover the underlying linguistic patterns.

Jang et al (2022) proposed the Affinity Prober as a specialized probing mechanism to investigate token relationships in self-attention-based language models. Their research applied the Affinity Prober algorithm to analyze how the BERT-base-cased model interprets well-formed and ill-formed sentences. According to Jang et al (2022), the decoding of token relationships extracted from attention scores, known as Lexical Categories, revealed noteworthy patterns in syntactic linguistic phenomena across different layers in the BLiMP benchmark (Warstadt et al., 2020). These patterns were observed in both well-formed and ill-formed sentences, providing valuable insights into the nature of syntactic processing within the model. Conversely, semantic linguistic phenomena displayed similar patterns. Furthermore, upon closer examination of specific phenomena such as wh-questions and negative polarity items (NPI) using the Affinity Prober, noteworthy distinctions in token relationships became evident. These distinctions provide valuable insights into the intricate workings of the model's syntactic processing when confronted with these linguistic constructs. Specifically, the study brought attention to distinct discrepancies in token relationships between well-formed and ill-formed sentences, particularly in the context of wh-questions.

This study aims to reinforce the methodology proposed by Jang et al (2022) through additional experiments. We begin by providing an overview of the research methodologies employed in related studies in Section 2, emphasizing the distinctiveness and significance of our research approach. In Section 3, we present experimental evidence to demonstrate that BERT autonomously learns linguistic knowledge related to part-of-speech by leveraging lexical categories. To achieve this, we introduce the ADTRAS algorithm (An Algorithm for Decrypting Token Relationships within Attention Scores) and combine it with lexical categories to analyze patterns at each layer of BERT. Our experiments focus on comparing the patterns observed in BERT when it is fine-tuned on specific tasks in the GLUE and SuperGLUE datasets and when it is not fine-tuned. We show the importance of BERT's part-of-speech processing and report on the general tendencies of layers that concentrate on content words and function words.

In Section 4, we shift our attention to the core of our study, introducing experiments using the Affinity Prober to analyze patterns in syntactic linguistic phenomena processed by BERT. We revisit Jang's (2022) research to explain our decision to focus solely on syntactic linguistic phenomena. We redefine the algorithm of the Affinity Prober, provide a clearer explanation of Affinity Relationship and Affinity Ratio. We then introduce the BLiMP dataset consisting of minimal pairs and the linguistic phenomena it covers. To facilitate comprehensive analysis, we present the Classification of Sentence Sequencing (CSS) as a Finetuning Strategy that indirectly learns from minimal pairs.

In the results section, our focus shifts towards understanding how BERT interprets linguistic phenomena in a fine-tuned setting, employing the Affinity Prober. By closely analyzing the patterns exhibited by BERT during the processing of sentences in various linguistic phenomena, we categorize similar patterns based on this information. Additionally, by establishing connections between the observed layer tendencies using the ADTRAS algorithm, we aim to generalize the behavior of BERT layers when processing sentences with syntactic phenomena, following the CSS approach as the fine-tuning strategy.

Finally, in Chapter 5, we provide a summary of our research contributions and discuss the limitations of our study, offering insights into future directions for research.

The key contributions of our study are as follows:

1. Proposal of ADTRAS Algorithm: The ADTRAS algorithm is introduced to analyze patterns at each layer of BERT, strengthening Jang's (2022) methodology and enhancing the interpretability of token relationships within BERT's attention scores. Our algorithm successfully captures significant linguistic movements within attention scores. Can we observe any explainable patterns in the activated neurons of continuous prompts through layers?
2. Experimental Evidence on BERT's Part-of-Speech Learning: Through empirical experiments, we demonstrate that BERT autonomously learns language knowledge related to part-of-speech by utilizing lexical categories. This finding supports the notion that BERT possesses an inherent understanding of grammatical categories.
3. Insight into Layer Tendencies: We provide insights into the general

tendencies of BERT's layers when processing content words and function words. By analyzing patterns at each layer, we uncover BERT's processing characteristics associated with different word types.

4. Introduction of Classification of Sentence Sequencing (CSS): We introduce CSS as a Finetuning Strategy, enabling indirect learning from minimal pairs. CSS facilitates a more comprehensive analysis of the relationship between minimal pairs and the underlying linguistic phenomena, leading to deeper insights into BERT's interpretation of linguistic patterns.
5. Examination of Syntactic Linguistic Phenomena: Using the Affinity Prober, we explore the patterns exhibited by BERT in processing syntactic linguistic phenomena. The analysis focuses on specific phenomena using the BLiMP dataset, highlighting the potential of the Affinity Prober in understanding syntactic structures processed by BERT.
6. Clustering of Similar Linguistic Phenomena: Through the Affinity Prober's analysis, we trace patterns exhibited by BERT layers and cluster similar linguistic phenomena, enabling a better understanding of their interrelationships.

Chapter 2. Related Works

This chapter offers a comprehensive review of significant research examining the linguistic knowledge inherent in language models, with a specific emphasis on BERT. The chapter is segmented into three sections: Section 2.1 elucidates the Probing Classifier Framework and its role in syntactic analysis; Section 2.2 dives into the exploration of syntactic tree generation in correlation with neural networks; and finally, Section 2.3 reviews studies that delve into the intricate relationship between BERT and linguistics.

2.1. Unveiling Linguistic Insights: The Probing Classifier Framework

Before the emergence of Transformers, researchers extensively explored the syntactic analysis in context-based representations. Among the analytical methods, Probing Classifiers emerged as a viable means of studying the syntactic nuances of neural network models in the natural language processing (NLP) realm. Noteworthy contributions include those from Belinkov (2017), who examined how Neural Machine Translation (NMT) architecture comprehends word structure and part-of-speech (POS). Blevins et al. (2018) posited that RNN models trained on diverse NLP tasks could induce syntactic hierarchy without explicit guidance. Furthering this field, Conneau et al. (2018) put forward ten probing tasks for assessing linguistic properties, while Hupkes et al. (2018) utilized Diagnostic Classifiers, a supervised method, to investigate how RNN models interpret syntactic hierarchy. Hewitt and Manning (2018), recognizing the limited explanatory capabilities of neural network models in revealing parse trees within deeply learned contextual models, proposed a structural probe. They asserted that ELMo and BERT exhibit robust syntax based on minimum spanning trees. Yet, the Probing Classifier Framework is not without its limitations; Belinkov (2022) highlighted the ambiguity in the choice of classifier for diverse contexts.

2.2. The Interplay of Syntactic Tree and Neural Networks

One of the crucial research areas in extracting implicit linguistic knowledge within neural networks revolves around the generation of syntactic tree structures. A long-standing challenge in NLP has been to induce such structures in an unsupervised manner. Pioneering contributions from Klein and Manning (2001; 2002; 2004) implemented probabilistic part-of-speech tagging based on treebank sequences, laying the foundation for unsupervised parsing utilizing phrase-structure grammar and tree-based models. The emergence of deep learning, as emphasized by LeCun et al. (2015), and the introduction of RNN models by Hochreiter and Schmidhuber (1997), brought significant attention to the field and propelled extensive research efforts in unsupervised syntactic structure induction based on RNN models. The advent of the Transformer architecture directed the research on syntactic structures beyond the design of neural network models strictly for inducing these structures. Syntax-BERT (Bai et al., 2019) proposed syntactic attention layers by inducing MASKs based on constituency trees (Chen and Manning, 2014) and dependency trees (Zhu et al., 2013). Li et al. (2020) further refined this process by devising a Mask Matrix based on dependency parsing information, integrating it into BERT's attention scores to enhance its performance.

2.3. BERT and Linguistics

The Bidirectional Encoder Representation from Transformers (BERT) model, introduced by Devlin et al. (2019), has made remarkable strides in the field of NLP. BERT is a transformer-based language model that leverages the power of self-attention mechanisms to encode bidirectional contextual information, allowing it to achieve state-of-the-art performance on various NLP tasks.

BERT's architecture is rooted in the transformer model proposed by Vaswani et al. (2017), which introduced the concept of self-attention mechanism, enabling efficient parallel processing of tokens in a sequence. This mechanism allows BERT to capture the contextual information for each token, making it inherently bidirectional and resolving some of the limitations of previous unidirectional models. Pre-training is a crucial aspect of BERT model. During pre-training, BERT is exposed to large corpora and learns contextual representations by predicting masked words in a sentence (Masked Language Modeling, MLM) and predicting whether two sentences follow each other (Next Sentence Prediction, NSP). This pre-training process enables BERT to develop a deep understanding of language structures and relationships, which can be further fine-tuned for specific downstream tasks. By fine-tuning BERT, it is adapted to various NLP tasks such as text classification, named entity recognition, question-answering, etc. In this process, BERT's pre-trained representations are combined with task-specific classifier layers and fine-tuned on smaller specific datasets. This fine-tuning strategy allows BERT to transfer its knowledge learned during pre-training to new tasks effectively.

BERT's remarkable performance across various NLP tasks, particularly linguistic tasks, has generated significant interest, leading to extensive explorations into its encoding and decoding mechanisms for linguistic information. Numerous studies have probed the relationship between BERT and linguistics (Rogers et al., 2021), with this section specifically concentrating on studies most relevant to our research. Jawahar et al. (2019) have explored BERT's capabilities in capturing structural information in language. Their investigation reveals that different layers of BERT are dedicated to encoding specific linguistic features. Lower layers tend to focus on phrase-level information, middle layers concentrate on syntactic aspects, while top layers emphasize semantic understanding. This demonstrates BERT's ability to effectively represent different levels of linguistic structures.

Contrarily, Htut et al. (2019) conducted fine-tuning experiments on syntax-oriented and semantics-oriented datasets to identify significant shifts in attention weights and to extract dependency relations. They try to understand the changes in BERT's attention weights following fine-tuning on two distinct datasets: one syntax-oriented (CoLA) and the other semantics-oriented (MNLI). Although their findings indicate attention heads tracking individual dependency types, the generalization of such learned representations is limited, shedding light on the challenges in adapting BERT's attention mechanisms to different tasks. Although they found BERT's attention heads tracked individual dependency types, they noted this might not be a universal trait.

Contrasting these findings, Kovaleva et al. (2019) reported an absence of significant attention shifts in BERT, postulating that attention maps might be influenced more by pre-training tasks than by task-specific linguistic reasoning. Their research primarily investigated whether BERT's fine-tuning on a specific task leads to self-attention patterns that emphasize particular linguistic features.

Chapter 3. Generalization of Layer-Wise Attention Using ADTRAS Algorithm

In this chapter, we experimentally demonstrate that BERT learns linguistic knowledge about lexical categories during the fine-tuning process and reveal that this knowledge can be generalized to explain the properties of BERT layers in terms of categories. To conduct our experiments, we propose the ADTRAS (An Algorithm for Decrypting Token Relationships within Attention Scores) algorithm. We train the BERT-base-cased model on six tasks from the GLUE benchmark and examine the attention shift in BERT before and after fine-tuning using the ADTRAS algorithm. Ultimately, we uncover the existence of distinct properties within each layer of BERT and suggest the potential for layer generalization. Our findings offer valuable insights into the possibility of generalizing the behavior and characteristics of BERT layers.

3.1. Binary Categorization of Part-of-Speech in Sentences: Content Words and Function Words

In this experiment and for further experiment in Section 4, following Carpenter (1983), the part-of-speech information within sentences was binary-categorized as content words and function words. The part-of-speech information needed for this categorization was obtained through the NLTK (Natural Language Toolkit) module.[1]

[1]NLTK Module: https://github.com/nltk/nltk

- function words = {"CC", "MD", "DT", "EX", "IN", "PDT", "POS", "TO", "WDT", "WP", "WP\$", "WRB", "RP"}
- content words = {"NN”, “NNS”, “NNP”, “NNPS”, “CD”, “FW”, “JJ”, “JJR”, “JJS”, “PRP”, “PRP\$”, “RB”, “RBR”, “RBS”, “VB”, “VBD”, “VBG”, “VBP”, “VBZ”, “VBN”, “UH”}

The function words include coordinating conjunctions, modal verbs, determiners, existential 'there', prepositions and subordinating conjunctions, predeterminers, possessive endings, infinitive 'to', wh-determiners, wh-pronouns,

NLTK TAG	Description
CC	Coordinating Conjunction
MD	Modal
DT	Determiner
EX	Existential There
IN	Preposition, Subordinating Conjunction
PDT	Pre-determiner
POS	Possessive Ending
TO	infinitive to
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive Wh-pronoun
WRB	Wh-adverb
RP	Particle

Table 3.1: Description of NLTK Part-of-Speech Tags on function words

possessive wh-pronouns, wh-adverbs, and particles.

NLTK TAG	Description
NN	Noun (singular)
NNS	Noun (plural)
NNP	Proper Noun (singular)
NNPS	Proper Noun (plural)
CD	Cardinal Digit
FW	Foreign Word
JJ	Adjective
JJR	Adjective (comparative)
JJS	Adjective (superlative)
PRP	Personal Pronoun
PRP$	Possessive Pronoun
RB	Adverb
RBR	Adverb (comparative)
RBS	Adverb (superlative)
VB	Verb (base form)
VBD	Verb (past form)
VBG	Verb (gerund, present participle)
VBP	Verb (singular, present, non 3rd person)
VBZ	Verb (singular, present, 3rd person)
VBN	Verb (past participle)
UH	Interjection

Table 3.2: Description of NLTK Part-of-Speech Tags on content words

In contrast, the content words include nouns, plural nouns, singular proper nouns, plural proper nouns, cardinal numbers, foreign words, adjectives, comparative adjectives, superlative adjectives, personal pronouns, possessive pronouns, adverbs, comparative adverbs, superlative adverbs, base form verbs, past tense verbs, gerunds or present participle verbs, present tense verbs (non-3rd person singular), present tense verbs (3rd person singular), past participles, and interjections.

3.2. ADTRAS Algorithm

Our primary objective in this chapter is to investigate the linguistic characteristics and attention shift within the layers of BERT, with a specific emphasis on shifts in probabilistic scores within BERT's attention matrix. To accomplish this, we introduce the ADTRAS (An Algorithm for Decrypting Token Relationships within Attention Scores) algorithm, which allows for the decryption of token relationships while preserving the original attention values. ADTRAS is designed to work with multi-layered models like BERT and aims to uncover the connections between tokens that carry significant weights in the attention scores. Our main focus is to comprehend the relational structure of tokens, particularly in terms of lexical categories or Part-of-Speech. Additionally, the ADTRAS algorithm facilitates the extraction and understanding of syntactic configurations, semantic relationships between words, and causal correlations.

In the context of utilizing ADTRAS with BERT and analyzing lexical categories, our procedure begins by tokenizing and formatting the input sentence using a BERT model, represented as M. This step includes the incorporation of special tokens like *CLS* and *SEP* to ensure compatibility with the BERT model. Subsequently, the algorithm obtains the self-attention weights across all layers, denoted as A, from M

and calculates the mean across the heads in each layer, denoted as $\bar{A}$. Our analysis primarily focuses on meaningful tokens, excluding special tokens such as *CLS* and *SEP*. This process can be represented as:

$$\bar{A}_m = ExcludeSpecialTokens(\bar{A})$$

If the words are segmented into sub-tokens during tokenization, the attention weights are averaged by combining sub-tokens, denoted as

$$\bar{A}_{avg} = AverageSubtokenWeights(\bar{A}_m)$$

For each token, the algorithm identifies the token with the highest attention score max_{score} from the sequence of tokens E_T which contains t number of tokens $\{E_1, \ldots, E_t\}$. In cases where a token's attention is predominantly self-directed, the algorithm selects the second-highest attention score. This selection process is represented as $max_{idx} = argmax(E_T)$, and if $E_T = top_1$ itself , then $\max_{idx} = \mathrm{argmax}(E_T \setminus top_1)$, where $E_T \in \bar{A}_{avg}$.

The selected tokens are then assigned to their corresponding pre-determined lexical categories. Subsequently, the algorithm updates the frequency count for each lexical category.

In conclusion, the relative attention ratio for each lexical category is computed by normalizing the frequency count of each category by the total frequency count of all the different lexical categories, thus alleviating biases. Mathematically, this can be represented as

$$R_l = \frac{f_l[lexcat]}{\sum f_l[lexcat]}$$

By deriving the attention ratios R, which could explain the relationship between tokens in a sentence across all layers, we can perform layer-wise analysis using the ADTRAS algorithm. This enables us to examine the distribution patterns of attention within each layer. The summarized steps are provided in Alg 1.

Algorithm 1 ADTRAS

```
function ADTRAS(x)
if pair of sentences {x_1, x_2} ∈ X then
    E_T ← Embedding(cls, sep, x_1, x_2)
else
    E_T ← Embedding(cls, sep, x_1)
end if
A ← Attention(E_T)
Ā ← mean(A)
Ā_m ← ExcludeSpecialTokens(Ā)
Ā_avg ← AverageSubtokenWeights(Ā_m)
for l in layers do
    for E_T in Ā_avg do
        max_idx ← argmax(E_T)
        if max_idx == E_T then
            max_idx ← argmax(E_T \ max_idx)
        end if
        lexcat ← MapCategory(max_idx), where {Con., Fun.} ∈ lexcat
        f_l[lexcat] ← f_l[lexcat] + 1
    end for
    R_l ← f_l[lexcat] / ∑(f_l[lexcat])
    return R
end for
∑_{k=1}^{l} R_l
end function
```

3.3 The General Language Understanding Evaluation (GLUE)

In this study, we conducted an experiment using the BERT-base-cased model and focused on the tasks from the GLUE benchmark (Wang et al., 2018; Wang et al., 2019). Our goal was to fine-tune the model on a diverse range of tasks that require different types of semantic or syntactic information. Specifically, we selected six tasks that cover a wide spectrum of linguistic aspects.

3.3.1. The Corpus of Linguistic Acceptability (CoLA)

The Corpus of Linguistic Acceptability (CoLA) dataset, introduced by Warstadt et al (2018), is a benchmark in Natural Language Processing (NLP) that assesses models' ability to determine the grammatical acceptability of English sentences. Comprising 10,657 English sentences from various linguistic sources, the CoLA dataset is annotated to distinguish between grammatically acceptable and unacceptable instances. It focuses on making binary predictions about the grammatical acceptability of input sentences. The dataset presents challenges due to the disparity between grammatical acceptability and sentence meaning, which are often addressed during pre-training of NLP models. CoLA is an essential component of the GLUE benchmark, which evaluates the performance of different NLP models across various natural language understanding tasks.

3.3.2. The Microsoft Research Paraphrase Corpus (MRPC)

The Microsoft Research Paraphrase Corpus (MRPC) is a crucial task in NLP that assesses models' ability to determine the paraphrastic relationship between sentence pairs. Introduced by Dolan and Brockett in 2005, the MRPC dataset contains approximately 5800 sentence pairs sourced from web-based news content. Human annotators labeled each pair to indicate whether they exhibit paraphrastic properties. The MRPC task revolves around accurately categorizing sentence pairs as paraphrases or non-paraphrases. It is commonly approached as a binary classification problem, where models predict '1' for paraphrase pairs and '0' for non-paraphrase pairs. MRPC is part of the GLUE benchmark and evaluates models' comprehension of syntactic and semantic aspects, as well as their ability to recognize and generate paraphrases.

3.3.3. The Stanford Sentiment Treebank 2.0 (SST-2)

The Stanford Sentiment Treebank 2.0 (SST-2) is a dataset designed for sentiment analysis in NLP. Developed by Socher et al. in 2013, it builds upon the original Stanford Sentiment Treebank. With 67,349 English sentences extracted from movie review excerpts, the SST-2 dataset labels each sentence as positive or negative sentiment. It focuses on binary sentiment classification, removing neutral instances for simplicity and effective model training and evaluation. The SST-2 task aims to accurately determine the sentiment expressed in a given sentence, providing a testing ground for models' understanding of sentiment in text. SST-2 is part of the GLUE benchmark and enables evaluation and benchmarking of NLP models' performance across various natural language understanding tasks.

3.3.4. The Quora Question Pairs (QQP)

The Quora Question Pairs (QQP) dataset is a significant benchmark for evaluating NLP models' ability to identify semantically equivalent questions. Created by Quora to consolidate duplicate questions, the QQP dataset consists of over 400,000 question pairs. The task involves determining whether a pair of questions are duplicates or not, making it a binary classification problem. The QQP dataset presents challenges due to the variation in expressions used to ask essentially the same question. Models must understand the underlying semantic content of questions rather than relying solely on lexical matches.

3.3.5. The Multi-Genre Natural Language Inference (MNLI)

The Multi-Genre Natural Language Inference (MNLI) task evaluates NLP models' ability to identify semantic relationships between sentence pairs. Introduced by Williams et al. in 2017, the MNLI dataset contains approximately 433,000 sentence pairs, each labeled with textual entailment information. The pairs consist of a premise and a hypothesis sentence, and the task is to determine whether the premise entails, contradicts, or is neutral to the hypothesis. MNLI draws sentences from ten genres of written and spoken English, providing a diverse range of linguistic styles and lexical choices for evaluation. MNLI is included in the GLUE benchmark and serves as a rigorous evaluation of models' understanding of textual entailment and semantic relationships between sentences.

3.3.6. The Words in Context (WiC)

The Words in Context (WiC) task, part of the SuperGLUE evaluation suite, focuses on word sense disambiguation in NLP. Introduced by Wang et al. in 2019, the WiC task tests models' ability to determine the correct sense of a target word in two different contexts. The dataset provides pairs of sentences, each containing a target word, and models must determine whether the word has the same sense in both sentences. The WiC dataset consists of approximately 1,000 instances, labeled as 'True' if the target word retains the same sense and 'False' if the senses differ. This binary classification task requires a deep understanding of language and context beyond syntactic comprehension. The WiC task originated from the Word in Context dataset and provides a challenging evaluation for NLP models.

3.3.7. Summary

For each task, we fine-tune the bert-base-cased model. Additionally, we employ the ADTRAS algorithm to decode word attention relations, allowing us to identify notable shifts when examining the data through the lens of lexical categories. To classify and tag content words and function words, we utilize the NLTK (Natural Language Toolkit) module, following the definition provided by Carpenter (1983).

By conducting experiments on these diverse tasks and analyzing attention relations with respect to lexical categories, we aim to gain insights into the model's understanding and representation of semantic and syntactic information across different linguistic phenomena.

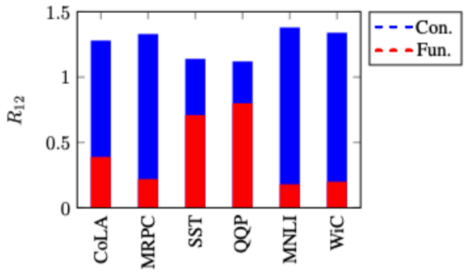
3.4 Evaluating Attention Variations in Lexical Categories on NLU tasks

In the results, we evaluate the six models on six distinct test datasets, both before and after fine-tuning. Using the ADTRAS algorithm, we analyze the changes in attention within the lexical category at each layer. This analysis allows us to examine the variations in attention patterns for different models and layers.

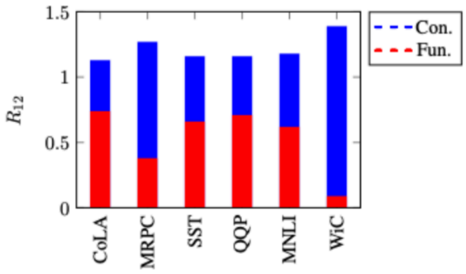
	Pretrained		**Finetuned**	
	Con.	Fun.	Con.	Fun.
CoLA	1.27	.38	1.12	**.73** (+.35)
MRPC	1.32	.21	1.26	**.37** (+.16)
SST	1.13	.70	**1.15**	.65 (-.05)
QQP	1.11	.79	**1.15**	.70 (-.09)
MNLI	1.37	.17	1.17	**.61** (+.44)
WiC	1.33	.19	**1.38**	.08 (-.08)

Table 3.3: Changes in Attention Distribution Across Lexical Categories from Pre-trained Model to Fine-tuned Model

3.4.1 Intrinsic Learning of Lexical Categories in BERT for Downstream Task



(a) Pretrained Model



(b) Finetuned Model

Figure 3.1: Changes in Attention Distribution Across Lexical Categories from Pre-trained Model to Fine-tuned Model

This section explores the impact of fine-tuning BERT on attention weights across various downstream tasks, offering valuable insights into the learning capabilities of self-attention in relation to lexical categories. Specifically, our analysis focuses on the last layer of BERT, which previous studies (Liu et al., 2019; Kovaleva et al., 2019; Hao et al., 2019) have identified as task-specific. The findings highlight significant attention shifts dependent on the task type, as depicted in Figure 3.1 and summarized in Table3.3.

For example, when fine-tuning BERT for the CoLA and MRPC tasks, which emphasizes syntactic structures, we observe an increase in attention towards function words and a decrease in attention towards content words. On the other hand, fine-tuning for the WiC task, which focuses on relationships among content words, leads to an increase in attention towards content words and a decrease for function words. This shift is intriguing because the fine-tuned model exhibits even higher attention to content words, surpassing the significant attention already present in the pretrained model. Moreover, tasks like SST-2 and QQP, prioritizing semantic aspects over syntactic ones, demonstrate an escalation in attention towards content words. In contrast, the MNLI task, which requires both syntactic and semantic understanding, exhibits a substantial amplification in attention towards function words. These observations indicate a strong connection between the MNLI task and the utilization of syntactic information.

To summarize, tasks involving syntactic information (CoLA, MRPC, MNLI) show increased attention weights on function words, while tasks emphasizing semantic information (SST-2, QQP, WiC) exhibit heightened attention on content words (refer to Table 3.3). These findings suggest that as language models undergo fine-tuning for specific objectives, they acquire inherent linguistic knowledge related to lexical categories.

	Con.			**Fun.**		
	top_1	top_2	top_3	top_1	top_2	top_3
CoLA	L12	L1	L11	L2	L8	L4
MRPC	L11	L12	L1	L8	L2	L9
SST	L1	L11	L12	L8	L2	L4
QQP	L1	L11	L12	L8	L9	L4
MNLI	L12	L11	L1	L8	L2	L4
WiC	L11	L12	L10	L2	L8	L4

Table 3.4: Top 3 Layers which mostly attend on the content words and function words on 6 downstream tasks

3.4.2 Generalization of Layer-Wise Attention in Fine-Tuned BERT Models

Table 3.4 provides a comprehensive summary of the top three layers in each fine-tuned model, highlighting their highest attention to content words and function words. Interestingly, despite the variations in the fine-tuning process for each model, we can still observe consistent linguistic patterns in relation to lexical categories. The results in Table 3.4 demonstrate that Layers 1, 10, 11, and 12 predominantly focus on content words, while Layers 2, 4, 8, and 9 primarily focus on function words. This finding challenges previous studies that suggested BERT layers cannot be generalized (Htut et al., 2019; Kovaleva et al., 2019). Through the application of the ADTRAS algorithm, we successfully generalize the linguistic characteristics of BERT layers across six different downstream tasks.

Chapter 4. Probing Intrinsic Linguistic Knowledges of Deep Learning-based Language Model using Affinity Prober

In Section 3, we observed noteworthy changes in attention scores using the ADTRAS algorithm during the fine-tuning of BERT. This algorithm, focused on the Lexical Category, revealed a tendency to prioritize the relevant lexical categories based on the specific task objectives. One intriguing finding was the identification of layers within each of the six fine-tuned models that exhibited distinct attention patterns towards content and function words. This indicates the ability to fine-tune BERT to pay closer attention to specific linguistic aspects, tailored to the objectives of each experiment.

The purpose of this section is to explore the relationships between different layers of BERT across various syntactic language phenomena, specifically focusing on lexical categories. This investigation is motivated by the belief that BERT possesses inherent linguistic knowledge in relation to lexical categories. Our focus is specifically on syntactic language phenomena, based on the evidence presented in Jang's 2022 study. This study revealed meaningful differences between well-formed and ill-formed sentences in terms of syntactic language phenomena, as analyzed from the perspective of lexical categories. Such distinctions were not observed in semantic language phenomena.

In this chapter, we begin by presenting the findings from Jang's (2022) study. We then proceed to refine and revisit the Affinity Prober algorithm. Additionally, we provide a concise overview of the syntactic language phenomena that will be utilized in our forthcoming experiment.

4.1 Jang et al (2022)

In Jang's (2022) study, a novel methodology called the Affinity Prober was introduced to investigate the decision boundaries of deep learning-based pre-trained language models when processing linguistic phenomena. The Affinity Prober leverages the attention scores of the language model's self attention mechanism to extract word affinity relationships, particularly focusing on the relationship between content and function words.

In the context of syntactic language phenomena, Jang discovered that the top layers of the language model exhibited decision boundaries that could explain the differences between well-formed and ill-formed sentences through lexical affinity. He also observed a strong reinforcement of the relationship between function words in syntactic language phenomena at the higher layers of BERT. Furthermore, Jang successfully delineated the acceptance decision boundary through the examination of wh-questions. However, he did not identify clear decision boundaries for distinguishing between well-formed and ill-formed sentences in semantic language phenomena. Ambiguity was commonly observed in the affinity relationship of minimal pairs involving negative polarity items. He found that semantic language phenomena prioritize relationships between content words, while little emphasis is placed on relationships between function words at all levels of BERT.

The Affinity Prober sets itself apart from existing probing methods by extracting universal language information from sentences in parallel. This distinction is significant. Moreover, Jang's study demonstrated the validity of the Affinity Prober by uncovering clear decision boundaries in the language model that revolve around lexical categories in syntactic language phenomena. By calculating the affinity relationship between content and function words, the study provided insights into how the bert-base-cased model interprets specific grammatical phenomena, particularly the distinction between declarative and non-declarative sentences. This further validated the usefulness of the proposed probing method based on pre-training-based language models.

4.2 Affinity Prober

In this section, we deeply examine the workings of the Affinity Prober and provide a more technical definition of its mathematical notation and Affinity Relationship. The Affinity Prober is a distinctive algorithm that utilizes attention scores to systematically extract and quantify the affinity relationships, represented as $<A, F>$, among words within a given context, specifically in Transformer-based pre-trained language models. The attention scores embed the semantic interconnections between words and serve as a robust foundation for identifying and characterizing these relationships.

4.2.1 Multi-Head Attention on Transformer Architecture

Self-attention, also known as scaled dot-product attention, forms the foundation. For a given set of query Q, key K, and value V matrices, the self-attention score is computed through a sequence of operations (Vaswani et al., 2017).

Firstly, the dot product of the query and key matrices is evaluated (QK^{T}), subsequently scaling the output by the square root of the dimensionality of the key vectors ($\sqrt{d_k}$). Following this operation, a softmax function is applied to these scaled scores, yielding a set of attention weights. These weights are multiplied with the value matrix V to yield the output of the self-attention mechanism. In formal mathematical terms, this sequence of operations is represented as:

$$Att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, *T* signifies the transposition of a matrix, and *softmax* is the softmax function.

$$softmax(x) = \frac{exp(x_i)}{\sum_i exp(x_i)}$$

Expanding on the self-attention mechanism, the multi-head attention paradigm allows the model to concentrate on various positions in parallel. Instead of implementing a singular attention function with one set of learned linear projections, the model performs *h* parallel attention functions, each with a different set of learned linear projections for the queries, keys, and values (Vaswani et al., 2017). Each attention function or 'head' i yields an output value, which are concatenated and linearly transformed to produce the final output. This can be formalized as:

$$MultiHead(Q, K, V) = Concat(h_1, \ldots, h_h)W_O,$$

where each

$$h_i = Att(Q \cdot W_{Qi}, K \cdot W_{Ki}, V \cdot W_{Vi})$$

In the above equations, W_{Qi}, W_{Ki}, W_{Vi} and W_O denote the model parameters to be learned, while *Concat* refers to the concatenation operation.

4.2.2 Affinity Relationship

Affinity Relationship (AR) represents a strong mutual correlation within a sentence, particularly between an "Affiner" and an "Affinee". Mathematically, if we consider *W* as the set of all words in a sentence and *Att(w)* as the attention score assigned to a specific word \$w\$, we can define the "Affiner" denoted as $A \in W$, and the "Affinee" denoted as $F \in W$, as follows:

- Affiner A: A word that maximizes the affinity score across the set of words.

$$A = argmax_{w \in W} Att(w)$$

- Affinee F: The word which receives the maximum attention score from the Affiner.

$$F = argmax_{w \in W} Att(A, w)$$

Here, *Att(A, w)* signifies the attention score assigned by A to word w. Hence, the "Affiner" is the word which assigns the highest attention score to another word F in the sentence, and this mutual relationship, expressed as $<A, F>$, is termed the Affinity Relationship. The Affinity Prober's approach to word interrelationships, thereby, provides a robust mathematical framework for exploring the associations within language models.

4.2.3 Probabilistic Distribution of Categorized Affinity Relationships

In the work conducted with the Affinity Prober, we position linguistic concepts as a foundation for word categorization, such as part-of-speech tagging. This paradigm enables an examination of the efficacy of pre-established linguistic concepts through their interactive behavior within the language model and facilitates the calculation of the affinity ratio between categories to study their respective impact on the model.

Take, for example, two categories *X* and *Y*, capable of serving as taxonomies for natural language. We can derive information about affinity relationships such as *<X, X>, <X, Y>, <Y, X>, and <Y, Y>*. Given that all words are binarized into their respective categories, the *<A, F>* relationships for all words can be parsed into four distinctive categories. This procedure leads us to the derivation of each affinity relationship's probability distribution, an attribute we define as the Affinity Ratio in equation below.

- Affinity Ratio: Suppose *C* denotes a categorization function mapping a word to a category (either *X* or *Y*). If \$N\$ represents the total number of words in a corpus and $N(c_1, c_2)$ is the count of *<A, F>* pairs where Affiner is categorized as c_1 and Affinee as c_2, the affinity ratio $AR(c_1, c_2)$ is formulated by:

$$AR(c_1, c_2) = N(c_1, c_2)/N, for\ c_1, c_2 \in X, Y$$

This equation expresses the probability distribution of the affinity relationships across categories *X* and *Y*.

4.2.4 The Algorithm of Affinity Prober on BERT

We represent a BERT model as *M*, which consists of *L* layers. Each layer *l* is equipped with *H* self-attention heads, resulting in a total of *L* × *H* self-attention operations. Specifically, the BERT-base model consists of 12 layers (*L=12*), with each layer containing 12 attention heads (*H=12*). Therefore, any given input sequence undergoes 144 (*L* × *H*) distinct self-attention operations.

During each self-attention operation, an attention score matrix is generated, capturing the semantic and syntactic correlations between tokens. Higher attention scores indicate stronger relationships, indicating that the model places greater emphasis on these token pairs when encoding the sequence.

The Affinity Prober algorithm is designed to interpret these attention scores as a measure of word "Affinity". For a given input sentence $s = \{w_1, w_2, ..., w_N\}$, the algorithm leverages the self-attention mechanism of M to establish Affinity Relationships for each word w_i. It identifies the word w_j that has the maximum attention score in relation to w_i across all layers and heads. This relationship, denoted as (w_i, w_j), is referred to as $AR(w_i)$ and can be expressed mathematically as:

$$AR(w_i) \ = \ argmax \ Att_l^h(w_i, w_j),$$

where $Att_l^h(w_i, w_j)$ is the attention score between w_i and w_j at layer l and head h. By applying this process to all words in s, we obtain a collection of Affinity Relationships that encompass the entire sentence, representing the word associations as perceived by the BERT model.

To investigate the layer-wise characteristics of BERT, we adapt the Affinity Prober to calculate the average attention head outputs for each layer. As a result, the equation is modified as:

$$AvgAtt_l(w_i, w_j) \ = \frac{1}{H}\sum_{h=1}^{H} Att_{l,h}\,(w_i, w_j),$$

which computes the average attention score across all heads in layer l between w_i and w_j. Then, the Affinity Relationship, computed with averaged attention across all heads at each layer, is given by:

$$AR_l(w_i) = argmax_{w_j} AvgAtt_l(w_i, w_j)$$

Here, $AR_l(w_i)$denotes w_j that shares the maximum averaged attention score with w_i at layer \$1\$. Consequently, w_j is recognized as the Affinee of w_i at layer *l*. This reformulation enables layer-wise operation of the Affinity Prober, averaging attention scores across all heads in a particular layer.

The Affinity Relationship extracted through the Affinity Prober focuses solely on strong connections between tokens. Leveraging the Affinity Prober opens up numerous research possibilities, such as precisely investigating the Dependency Parsing of sentence structures by tracking the relationships between specific words as they traverse through layers. Additionally, it is possible to map each token to specific linguistic concepts, such as parts-of-speech, using the Affinity Prober. This would enable tracking how the language model interprets parts-of-speech based on the relationships between them.

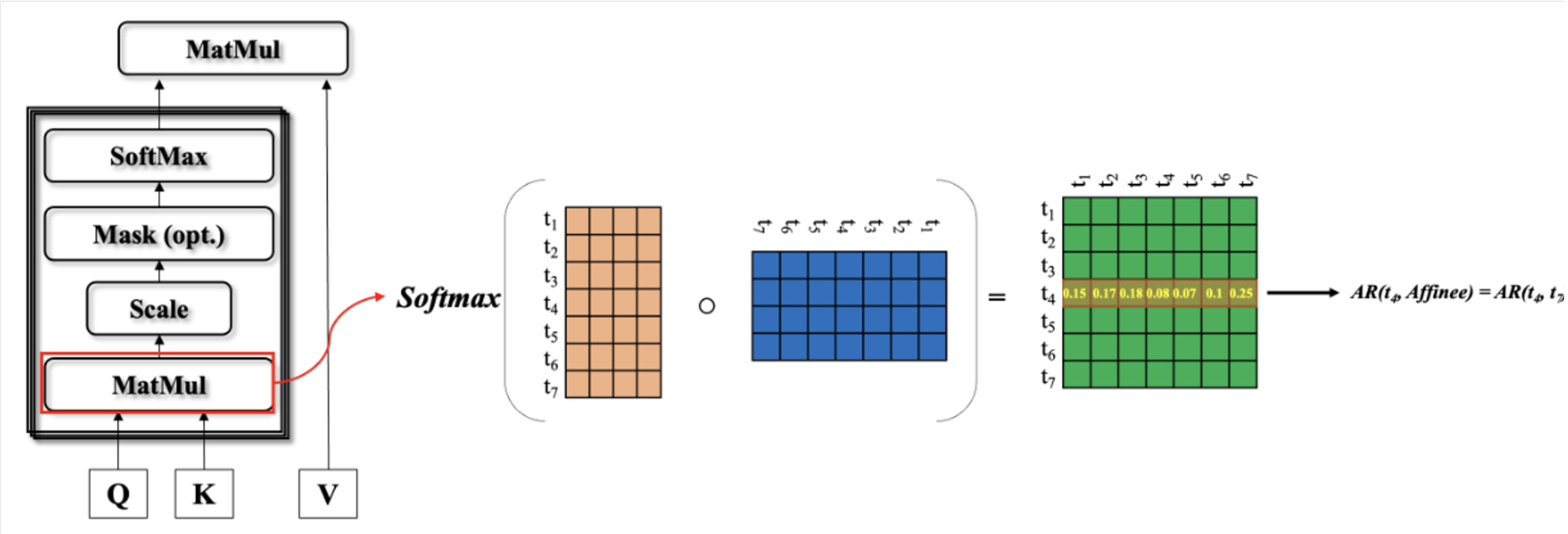


Figure 4.1: The process of probing tokens that have Affinity Relationship with Affiner t_4 using Affinity Prober. In this process, MatMul operation is performed for each layer's attention head, and the average of attention heads can also be used. In this paper, the goal is to extract the results of Affinity Relationship per each layer.

Chapter 5. The Benchmark of Linguistic Minimal Pairs (BLiMP)

The Benchmark of Linguistic Minimal Pairs (BLiMP), formulated by Warstadt et al (2020), serves as a rigorous evaluation benchmark to assess the linguistic understanding of language models. The dataset comprises 67 linguistic phenomena, each meticulously examined using a curated collection of 1000 minimal pairs.

Minimal pairs consist of two sentences that are nearly identical, except for one crucial difference. In these pairs, one sentence adheres to grammatical rules (well-formed), while the other violates them (ill-formed).[2]

The BLiMP dataset aims to test the ability of language models to distinguish grammatically correct sentences from flawed ones, focusing on subtle differences between them. It assumes that language models with strong linguistic knowledge gained from their training data should be capable of discerning such nuanced variations.

The dataset covers a wide range of linguistic phenomena, including agreement, case marking, filler-gap dependencies, and island effects, among others. The sentences in the dataset are intentionally kept simple, avoiding idiomatic or ambiguous structures to ensure a clear focus on the specific phenomena under examination.

Each phenomenon in the dataset is accompanied by a detailed description,

[2] i. a. well-formed sentence: The cat is sleeping on the bed.

b. ill-formed sentence. The cat is sleeps on the bed.

example sentences, and a concise discussion that explains the grammatical errors in the ill-formed sentences based on English grammar rules. This makes the dataset not only a valuable tool for evaluation but also a valuable resource for understanding the strengths and weaknesses of language models in acquiring different aspects of linguistic knowledge.

Due to the complexity of the BLiMP dataset as a benchmark, achieving a high score is a challenging task. Despite the simplicity of the sentences, the distinctions between well-formed and ill-formed sentences can be extremely subtle, requiring a deep understanding of English grammar for accurate classification.

5.1 Adjunct Island

Adjunct Island constraints are a type of the family of syntactic rules known as island constraints, which govern the circumstances under which a constituent can be moved from one position to another in a sentence, or whether it can not be moved at all. In general, an adjunct island refers to a syntactic configuration in which a word or phrase (usually a wh-word) is moved out of an adjunct clause, and this movement is typically considered to be unacceptable.

In the BLiMP dataset, the Adjunct Island tests would involve pairs of sentences where one violates the adjunct island constraint, and one does not.

(1) a. Who should Derek hug aftershocking Richard?

b. *Who should Derek hug Richard after shocking?

5.2 Animate Subject

5.2.1 Animate Subject Passive

The "Animate Subject Passive" is a category of grammatical phenomena which BLiMP includes for testing the capability of models to handle passive constructions in sentences where the subject is animate (i.e., a living entity).

In English, passive sentences are those where the subject is acted upon by the verb, and the agent of the action may be omitted or introduced by a prepositional phrase. In passive constructions, animate subjects typically receive an action rather than perform it. An example of such a sentence pair in BLiMP is following:

(2) a. The cat was chased by the dog.

b. *The cat was chased by the table.

In this pair, (2a) is grammatically correct and makes sense, as "the cat" (an animate entity) can logically be chased. (2b) is considered ungrammatical or nonsensical because semantically a table (an inanimate object) cannot chase a cat. Models successful on this task would need to understand the concept of animacy and its role in grammatical sentence construction.

5.2.2 Animate Subject Trans

The "Animate Subject Trans" subset in the BLiMP (Benchmark of Linguistic

Minimal Pairs) dataset pertains to instances of transitive syntactic constructions with animate subjects. A transitive construct necessitates the presence of both a subject and one or more objects.

Within the scope of the "Animate Subject Trans" classification, the emphasis is on the animate subject (a living entity) instigating an action that has a direct impact on an object. The BLiMP dataset presents pairs of sentences: one conforming to grammatical norms, and the other demonstrating an error. For example:

(3) a. The dog pursued the ball.

b. The dog pursued.

In the provided example, the first sentence (3a) abides by grammatical rules with "the dog" (an animate subject) executing the action (pursued) that directly involves an object ("the ball"). (3b) is grammatically incorrect due to the absence of an object for the transitive verb "pursued."

Models proficient in this specific task would be expected to grasp the concept of transitivity, as well as the requirement for animate subjects to be associated with an object in instances involving a transitive verb.

5.3 Causative

Causation entails a situation in which a specific action or event is instigated or facilitated by a causer. Within this context, the "causee" (the entity on which the action is performed) experiences a state change or action due to the actions of the "causer" (the agent initiating the action). We refer to instances extracted from the BLiMP dataset to elucidate this:

(4) a. Aaron breaks the glass.

b. *Aaron appeared the glass.

In the context of (4a), "Aaron" functions as the causer, effectuating the action of breaking, while "the glass" is the causee, undergoing the action. However, (4b) deviates from conventional causative use. Here, the verb "appeared" does not fit the traditional causative framework, leading to an ill-formed construction. In this setting, the verb "appeared" takes "Aaron" as a causer, which doesn't typically take. This example serves to underline the model's capability to distinguish well-formed and ill-formed causative sentences, thereby evaluating its understanding of causative phenomena.

5.4 Complex NP Island

The Complex Noun Phrase (NP) Island Constraint, also known as the Complex NP Constraint, is a syntactic rule that disallows extraction out of certain complex noun phrases.

In other words, it refers to a phenomenon where certain elements (such as a relative clause) within a complex noun phrase create a 'syntactic island'—an area of a sentence from which constituents cannot be moved or extracted, especially in questions and relative clauses. Consider the examples from the BLiMP dataset:

(5) a. Who aren't most hospitals that hadn't talked about most waitresses alarming?

b. *Who aren't most waitresses alarming most hospitals that hadn't talked about?

In this case, the NP "most hospitals" forms an island, which restricts the movement of constituents out of that island. The NP "most waitresses" is base-generated within the relative clause, and in (5a), it could not move out of the island created by "most hospitals." However, in (5b), "most waitresses" attempts to move across the island, which violates the island constraint and makes the sentence ungrammatical.

5.5 Coordinate Structure Constraint

The Coordinate Structure Constraint (CSC), an established axiom within linguistics, asserts that constituents such as words or phrases cannot be isolated from a single clause within coordinate structures (those combined by conjunctions such as "and" or "or").

5.5.1 Left Branch

An extrapolation of the principle above, known as the Coordinate Structure Constraint Complex Left Branch (CSC Complex Left Branch), stipulates a prohibition on extracting a constituent from the left (or initial) aspect of a coordinate structure that possesses complexity, such as subordination or embedding. To illustrate, consider a pair of exemplars from the BLiMP corpus:

(6) a. What senators was Alicia approaching and some teachers scaring?

b. What was Alicia approaching senators and some teachers scaring?

In (6a), "What senators" is the constituent extracted from the left branch of each clause in the coordinate structure: "Alicia was approaching [what senators]" and "some teachers scaring [what senators]". Each of these sentences could independently ask about the identity of the senators, and when combined with the conjunction "and", the sentence remains grammatically sound. Therefore, this sentence respects the CSC Complex Left Branch constraint.

Contrarily, in the case of (6b), "What" is extracted, and it is unclear to which part of the sentence it applies: "Alicia was approaching [what] senators" or " [what] some teachers scaring". Here, "what" is not tied to a specific constituent and its relation to the rest of the sentence is ambiguous. This ambiguity breaches the CSC Complex Left Branch constraint, rendering the sentence ungrammatical.

5.5.2 Object Extraction

The Coordinate Structure Constraint (CSC) "Object Extraction" paradigm entails the displacement of an object from one of the conjuncts in a coordinated structure to the sentence-initial position.

Extraction in linguistic parlance constitutes a mechanism wherein a lexical item, a phrase, or a clause is translocated from a larger structure, engendering a gap. This operation is most commonly associated with question formation, but it also surfaces in the creation of relative clauses and other syntactic constructions. Let us consider a pair of sentences from the BLiMP corpus:

(7) a. Who were all men and Eric leaving?

b. *Who were all men leaving and Eric?

In (7a), the pronoun "who" operates as the object of the action executed by the coordinated unit "all men and Eric". This sentence conforms to standard English grammar and is deemed well-formed as "who" serves as the object of the action carried out by the entire coordinated entity.

In constrast, in (7b), "who" is conceived as the object of the action executed solely by "all men". However, the Coordinate Structure Constraint forbids the extraction of "who" from a single conjunct ("all men") while leaving the remaining conjunct ("Eric") unrelated to the extracted object. Consequently, this sentence contravenes English syntactic norms, resulting in an ill-formed construction.

5.6 Drop Argument

"Drop Argument" in linguistics refers to a phenomenon where certain verbs allow for their arguments (subjects, objects, etc.) to be omitted or "dropped" without refers to the sentence ungrammatical.

Specifically, certain verbs, often called 'unergative verbs' such as 'run', 'sing', 'tour', are often found in contexts where the verb takes an agent as its subject without the complements as its object. For example, in the sentence "John is running", the verb 'run' does not require a direct object for the sentence to be grammatical.

However, not all verbs allow for their arguments to be dropped. These are often called 'transitive verbs', like 'reveal', 'find', 'hit', etc., which typically require a direct object. If the direct object is dropped, the sentence usually becomes ungrammatical. Let's consider the examples from the BLIMP dataset:

(8) a. Travis is touring.

b. *Travis is revealing.

In (8a), 'touring' is an unergative verb that doesn't require a direct object, so the sentence is grammatical even when the object is dropped. In contrast, in (8b), 'revealing' is a transitive verb which requires a direct object, so when the object is dropped, the sentence becomes ungrammatical.

5.7 Ellipsis N-bar

The syntactic phenomena of "N-bar Ellipsis" pertains to the construct wherein a fragment of an N-bar (a syntactic constituent typically encompassing an adjective and a noun) is subject to omission given its inferability from context.

The underlying principle of N-bar Ellipsis stipulates that constituents such as adjectives and nouns established in an antecedent portion of a sentence can be strategically omitted in a subsequent part, provided their contextual inference is preserved. Importantly, this presupposes a correspondence in syntactic structure and semantic content between the elided and the inferred elements. Consider the following instances derived from the BLiMP dataset:

(9) a. Dawn's ex-husband wasn't going to one rough grocery store and Becca wasn't going to many.

b. *Dawn's ex-husband wasn't going to one grocery store and Becca wasn't going to many rough.

In (9a), the phrase "rough grocery store" qualifies as an N-bar, with the term "rough grocery store" being validly elided in the second clause, given its implicit presence in the initial part of the sentence, thus referring the sentence syntactically well-formed.

Conversely, in (9b), the ellipsis of "grocery store" is syntactically flawed. This discrepancy stems from a structural mismatch between the elided component "many rough" and its antecedent in the sentence's initial clause, namely "grocery store". As such, the sentence contravenes English syntactic norms, and is deemed ill-formed.

5.8 Inchoative

Inchoative verbs represent a distinct class of verbs that manifest a transition in state. These verbs, rather than indicating an action instigated by the subject, instead signify a change being undergone by the subject. Consider the ensuing examples curated from the BLiMP corpus:

(10) a. Patricia had changed.

b. *Patricia had forgotten.

In instance (10a), the sentence "Patricia had changed" conforms to the grammatical rules, as "changed" is an inchoative verb that encapsulates a state transformation within the subject "Patricia".

On the other hand, sentence (10b) "*Patricia had forgotten.", employs the verb "forgotten" which does not conform to the inchoative verb schema as it fails to signify a change in state. Consequently, this sentence is deemed ill-formed within the context of inchoative verbs.

5.9 Intransitive

Intransitive predicates are those that do not necessitate a direct object to complete their semantic proposition, contrasting with transitive predicates that demand one or more object complements. Exemplary instances from the BLiMP corpus illustrate this phenomenon:

(11) a. Anna's grandmothers aren't benefiting.

b. *Anna's grandmothers aren't arguing about.

In instance (11a), the verb "benefiting" appropriately operates in an intransitive capacity, not necessitating an object for semantic completeness, yielding a well-structured statement.

Contrarily, (11b) constructs an ill-formed utterance in English syntax as the predicate "arguing about" inherently demands an object to convey a comprehensive semantic intent, thereby violating the premise of intransitive predicates.

5.10 Transitive

The phenomenon of transitivity pertains to the ability of a verb to necessitate an object for the completion of its meaning. In the English language, specific verbs like "buy" or "consume" are identified as transitive due to their syntactic and semantic demand for an object - the recipient of the action. Inspect the ensuing instances derived from the BLiMP dataset:

(12) a. This cousin of Theodore buys some mushroom.

b. *This cousin of Theodore wept some mushroom.

In (12a), the verb "buys" is employed transitively, encompassing "some mushroom" as its object, which results in a well-formed grammatical construction.

Conversely, in (12b), the verb "wept" is generally recognized as intransitive, hence it does not customarily admit an object. Consequently, the presence of "some mushroom" following "wept" engenders a syntactically ill-formed sentence, breaching the grammatical conventions of English.

5.11 Left Branch Island

5.11.1 Echo Question

"Left Branch Island Echo Question" pertains to a constraint in which wh-words, when serving as the leftmost branch of a constituent, cannot be extracted to form an echo question. Echo questions, in essence, are a type of interrogative wherein the speaker replicates part of a previous statement to request additional clarification. Consider the examples provided from the BLiMP dataset:

(13) a. Edward has returned to which customers?

b. *Which has Edward returned to customers?

In (13a), the wh-word "which" serves as the leftmost branch of the complement of the prepositional phrase "to which customers" and its placement adheres to the grammatical rules, resulting in a well-formed echo question.

However, in (13b), an attempt is made to extract "which" from the prepositional phrase and move it to the beginning of the sentence. This violates the Left Branch Island constraint, resulting in a sentence that is not syntactically well-formed in English. The structure of the sentence indicates an echo question, but it does not adhere to the acceptable syntactic pattern, leading to an ill-formed construct.

5.11.2 Simple Question

"Left Branch Island Simple Question" phenomenon refers to a syntactic constraint that prohibits the extraction of a determiner (like 'whose', 'which', 'what', etc.) from a noun phrase (NP) in wh-questions. This constraint refers to such extraction ungrammatical, marking the structure as a syntactic island -- a part of a sentence from which certain constituents cannot be moved or extracted. Take the provided examples from the BLiMP dataset:

(14) a. Whose museums had Dana alarmed?

b. *Whose had Dana alarmed museums?

In (14a), the wh-word "whose" correctly precedes and modifies the noun "museums". This sentence represents a grammatically well-formed English question, adhering to the accepted rules of English syntax.

On the other hand, in (14b), an attempt is made to extract the determiner "whose" from the noun phrase and place it at the sentence's beginning. This violates the Left Branch Island constraint and thus refers to the sentence ungrammatical. The ill-formed structure indicates that "whose" does not correctly modify the noun "museums", resulting in a syntactically flawed English question.

5.12 Passive

"Passive" phenomenon pertains to a syntactic structure where the subject of the sentence is the entity that the action is performed upon rather than the entity performing the action. This contrasts with active sentences, where the subject performs the action denoted by the verb. Consider the provided examples from the BLiMP dataset:

(15) a. Lucille's sisters are confused by Amy.

b. *Lucille's sisters are communicated by Amy.

In sentence (15a), "Lucille's sisters" are the subject and the entity upon which the action (confusing) is performed. "Amy," in this context, is the agent performing the action. The verb "confused" is correctly used in the passive voice, leading to a grammatically well-formed English sentence.

Distinctively, in sentence (15b), "communicated" is not typically used in the passive voice in English, particularly without an indirect object or a prepositional phrase to complete its meaning. Thus, the sentence is considered ill-formed according to standard English syntax. In other words, "Amy" cannot passively "communicate" Lucille's sisters, making this sentence a violation of the rules governing passive structures in English.

5.13 Sentential Subject Island

"Sentential Subject Island" phenomenon in linguistics pertains to the restrictions on the movement of constituents out of sentential subjects, a scenario often referred to as an 'island' for movement. That is, sentential subjects are syntactic constituents from which movement is generally prohibited, forming an 'island'. Consider the following examples from the BLiMP dataset:

(16) a. Who has the waitress's observing Christine bothered?

b. *Who has the waitress's observing bothered Christine?

In sentence (16a), the question word "who" is intended to be the object of the action "bothering". This sentence is grammatically correct and well-formed because "who" is not extracted from the sentential subject "the waitress's observing Christine".

However, in sentence (16b), "who" is intended to be the object of the action "observing by the waitress". This sentence is ungrammatical because extraction from a sentential subject is generally disallowed in English. Thus, attempting to extract "who" from "the waitress's observing" results in a violation of the Sentential Subject Island Constraint, and the sentence is considered ill-formed according to standard English syntax.

5.14 Wh Island

Wh-Island phenomenon in linguistics refers to a situation where a wh-word (like "who", "what", "when", "where", "why", etc.) cannot be extracted from a clause that is already introduced by another wh-word. This is considered an 'island' constraint and movement out of this 'island' is generally restricted. Consider the following examples drawn from the BLiMP dataset:

(17) a. Who have those men revealed they helped?

b. *Who have those men revealed who helped?

In sentence (17a), the wh-word "who" is appropriately extracted from a clause that is not introduced by another wh-word. Therefore, this sentence adheres to the grammatical rules and is well-formed.

However, in sentence (17b), an attempt is made to extract "who" from a clause that has been introduced by another wh-word ("who helped"). The clause "who helped who" creates an island, and the lower "who" cannot be extracted. This extraction violates the WH-Island Constraint, and thus, the sentence is considered ill-formed or ungrammatical according to the rules of English syntax. In accordance with the restrictions stipulated by the WH-Island phenomenon, a wh-word cannot be extracted from a clause that is already introduced by another wh-word.

5.15 Wh Questions

5.15.1 Object Gap

Wh-Question Object Gap phenomenon in linguistics relates to the positional constraint of WH-words, typically interrogative words, in object positions. A WH-word as an object in a sentence can create a 'gap', its original place before syntactic derivations. Consider the following examples from the BLiMP dataset:

(18) a. Joel discovered the vase that Patricia might take.

b. *Joel discovered what Patricia might take the vase.

In the well-formed sentence (18a), "the vase" is the object that Patricia might take. However, in sentence (18b), an attempt is made to transform the sentence into a WH-question by moving "the vase" to the front, replacing it with "what". The resulting sentence is not grammatically correct in English due to the absence of the 'gap' created in the object position of "take". This sentence violates the rule that, in WH-question formation, the Wh-word should correspond to the gap it leaves behind, which is not the case here. Thus, sentence (18b) provides an instance of an ill-formed WH-Question Object Gap phenomenon.

5.15.2 Subject Gap

The Wh-Question Subject Gap phenomenon in linguistics concerns the positional constraint of WH-words, typically interrogative words, in subject positions. A Wh-word used as a subject can create a 'gap' in the position where it would ordinarily be located before it is moved to the front of the sentence or clause during the question formation process. Consider the following examples from the BLiMP dataset:

(19) a. Brian had questioned an association that can astound Diana.

b.*Brian had questioned who an association can astound Diana.

In the grammatically correct sentence (19a), "an association" is the subject that can astound Diana. However, in sentence (19b), an attempt is made to convert the sentence into a WH-question by moving "an association" to the front and replacing it with "who". The resulting sentence is not grammatically acceptable in English due to the absence of the 'gap' created in the subject position. This sentence violates the rule that in WH-question formation, the WH-word must correspond to the gap it leaves behind, which is not the case in this context. Therefore, sentence (19b) serves as an instance of the ill-formed WH-Question Subject Gap phenomenon.

Chapter 6. Experiment

This paper's objective is to analyze layer-wise outcomes using the bert-base-cased language model. Our focus is on the syntactic linguistic aspects of the BLiMP benchmark. To achieve this, we utilize the Affinity Prober to obtain Affinity Relationships. Initially, we investigate the Affinity Ratio for each layer, specifically centered on the part-of-speech within each linguistic phenomenon. We analyze this at the lexical category level, which represents a higher category of the part-of-speech. Subsequently, we extract the Affinity Relationship from both correct and incorrect sentences across all layers and compare the disparities. Lastly, we extract the Affinity Relationship centered around the trigger token $w_{trigger}$, which is responsible for the incorrect sentences. Our goal is to assess whether there are distinctions in distinguishing between correct and incorrect sentences for each linguistic phenomenon. Building upon Jang's (2022) research findings, we anticipate significant variations in $AR(A, w_{trigger})$ between *IF* and *WF* sentences in terms of syntactic linguistic phenomena.

6.1 Finetuning Strategy

In our research, we propose a novel methodology called Classification of Sentence Sequencing (CSS) as an alternative to traditional binary classification approaches for grammaticality judgment. CSS enables the bert-base-cased model to distinguish between grammatically well-formed and ill-formed sentences by providing it with data from minimal pairs of sentences. Following the example of benchmarks such as

Question Answering, Natural Language Inference, and Word-in-Context, where pairs of sentences (S_1 and S_2) are inputted into the model, CSS introduces a combination of grammatically correct and incorrect sentences to improve the model's understanding.

The task of CSS involves determining the correct sequence of a well-formed (*WF*) and an ill-formed (*IF*) sentence within a minimal pair. For example, if the *WF* sentence is labeled as S_1 and the *IF* sentence as $_{S2}$, it corresponds to a boolean value of 'True'. Conversely, if the *IF* sentence is labeled as S_1 and the *WF* sentence as S_2, it returns a boolean value of 'False'. The model is trained using cross-entropy loss between the predicted labels (L_{pred}) and the actual labels (L), similar to a Logistic Regression model.

The CSS approach provides a significant advantage by enabling the model to effectively distinguish between two sentences with grammaticality determined by minimal pair tokens. If the model successfully accomplishes this task, it suggests that the language model has independently incorporated intrinsic linguistic knowledge. During training, we combined all datasets labeled in the syntactic domain within the BLiMP Benchmark. After randomly assembling the dataset, it was divided into training and testing sets in an 80:20 ratio, resulting in a model fine-tuned with syntactic knowledge.

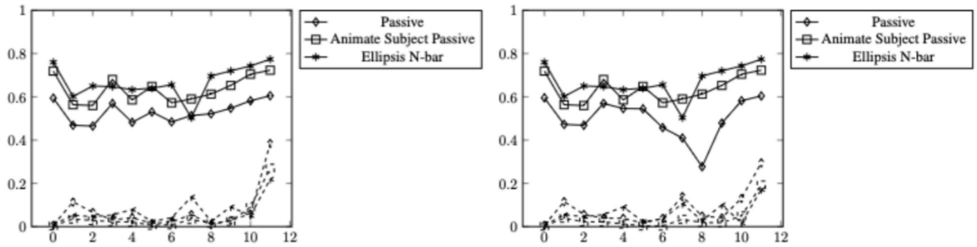
For training, we utilized the *Tanh* activation function and the cross-entropy loss function, along with the *AdamW* optimizer and a batch size of 16. The model underwent a total of 3 epochs of training. The training dataset consisted of 10,402 instances for the positive class (well-formed) and 10,398 instances for the negative class (ill-formed), while the test dataset included 2,095 positive instances and 2,065 negative instances which are extracted from the syntactic phenomena in BLiMP datasets. Impressively, our model achieved remarkable results, with a training performance of 99.9% accuracy and a loss value of 0.127. Equally impressive, the test performance mirrored these results with 99.9% accuracy and a loss value of 0.129. These findings support the effectiveness of our approach and its potential for high-impact applications.

The motivation behind adopting this specific training approach is to indirectly teach the model to discern the sequencing between well-formed and ill-formed sentences. However, it is important to note that the performance achieved through CSS training alone does not guarantee a complete distinction between the two categories.

6.2 Result: Clustering of Similar Linguistic Phenomena

In this section, we demonstrate how the Affinity Prober allows us to interpret the patterns obtained from the Affinity Relationship $AR(c_1, c_2)$ of each layer in BERT, where c belongs to the lexical category C defined in Section 3.1, based on different linguistic phenomena. We show that not all linguistic phenomena exhibit distinct patterns across layers in $AR(c_1, c_2)$. Instead, there are cases where similar patterns emerge, and these patterns can be grouped together. Through our observations, we are able to cluster these patterns into a total of four groups. This finding shows the interplay between linguistic phenomena and layer-wise patterns in the AR, ultimately enriching our understanding of language processing in BERT.

6.2.1 Group I: Passive and Ellipsis N-bar



(a) well-formed sentences (b) ill-formed sentences

Figure 6.1: The Affinity Relationship *(AR)* in three language phenomena associated with Group I: Passive, Animate Subject Passive, and Ellipsis N-bar. The solid lines in the figure represent the *AR(Con., Con.)*, while the dotted lines depict the *AR(Fun., Fun.)*.

The *AR(Con., Con)* results in Figure 6.1(a) demonstrate a consistent pattern across three datasets. All three exhibit a similar fluctuation range, marked by varying affinity ratio that fluctuate throughout the layers. An interesting commonality observed in all three datasets is an initial drop from the first to the second layer, indicating a uniform trend at the onset of the layers. If we interpret this drop in connection with the tendency of layers defined in section 3.4.2, the sharp drop of *AR(Con., Con.)* between Layer 1 and Layer 2 can be attributed to Layer 1's tendency to give high attention to content words, while Layer 2 shows a tendency to give high attention to function words. In other words, as the attention on function words increases in Layer 2, the relationship between content words relatively weakens.

Similarly, we can observe that the fluctuation in the middle layers and the maintenance of relationships between content words in the final layer align partially with the results presented in section 3.4.2. In the middle layers, which tend to focus on function words, the relationships between content words weaken again, only to be strengthened again around the final layer, where high attention is given to content words. An interesting point is that the relationship with function words also experiences a sudden drop in the final layer. This strengthening of function words in the final layer looks like a common phenomenon observed across all syntactic phenomena. This can be interpreted as a tendency that arises from the BERT model being fine-tuned in a CSS manner, where the task-specific considerations for distinguishing syntactic differences between two sentences are given to function words near the final layer.

When analyzing the linguistic phenomena of passive, animate subject passive, and ellipsis N-bar, there are some findings (Figure 6.2).

Passive The overall values of *AR(Con., Fun.)* are slightly lower for ill-formed sentences compared to well-formed sentences, indicating a slightly weaker feature in the context of ill-formed sentences. Although specific layers show small differences, such as lower values in the first and second layers for ill-formed sentences, the variations are not significant. However, there is a notable difference in layer 12, where the value of *AR(Fun., Con.)* is higher for ill-formed sentences, suggesting a slightly stronger feature in that layer for ill-formed sentences.

Animate Subject Passive Both well-formed and ill-formed sentences follow a similar data shape in *AR(Con., Fun.)* patterns. However, differences arise, such as a significantly higher value in the 9th layer of ill-formed sentences. Well-formed sentences display stability and a gradual decrease, while ill-formed sentences exhibit fluctuations and peaks. The 12th layer value is also higher for ill-formed sentences. In the *AR(Fun., Con.)* patterns, both types of sentences have a similar trend but diverge towards the end, with well-formed sentences declining more steeply. Ill-formed sentences fluctuate within a narrower range compared to well-formed sentences.

Ellipsis N-bar Both well-formed and ill-formed sentences show relatively high values in certain layers, indicating the presence of *AR(Con., Fun.)* in both cases. However, the values are slightly higher for well-formed sentences, suggesting a potentially stronger feature. Differences exist in specific layers, with some showing similar values while others exhibit notable differences. Similarly, in *AR(Fun., Con.)*, both types of sentences exhibit moderate values, but the overall values and specific layers differ. Ill-formed sentences have slightly higher values, indicating a potentially stronger feature in that context.

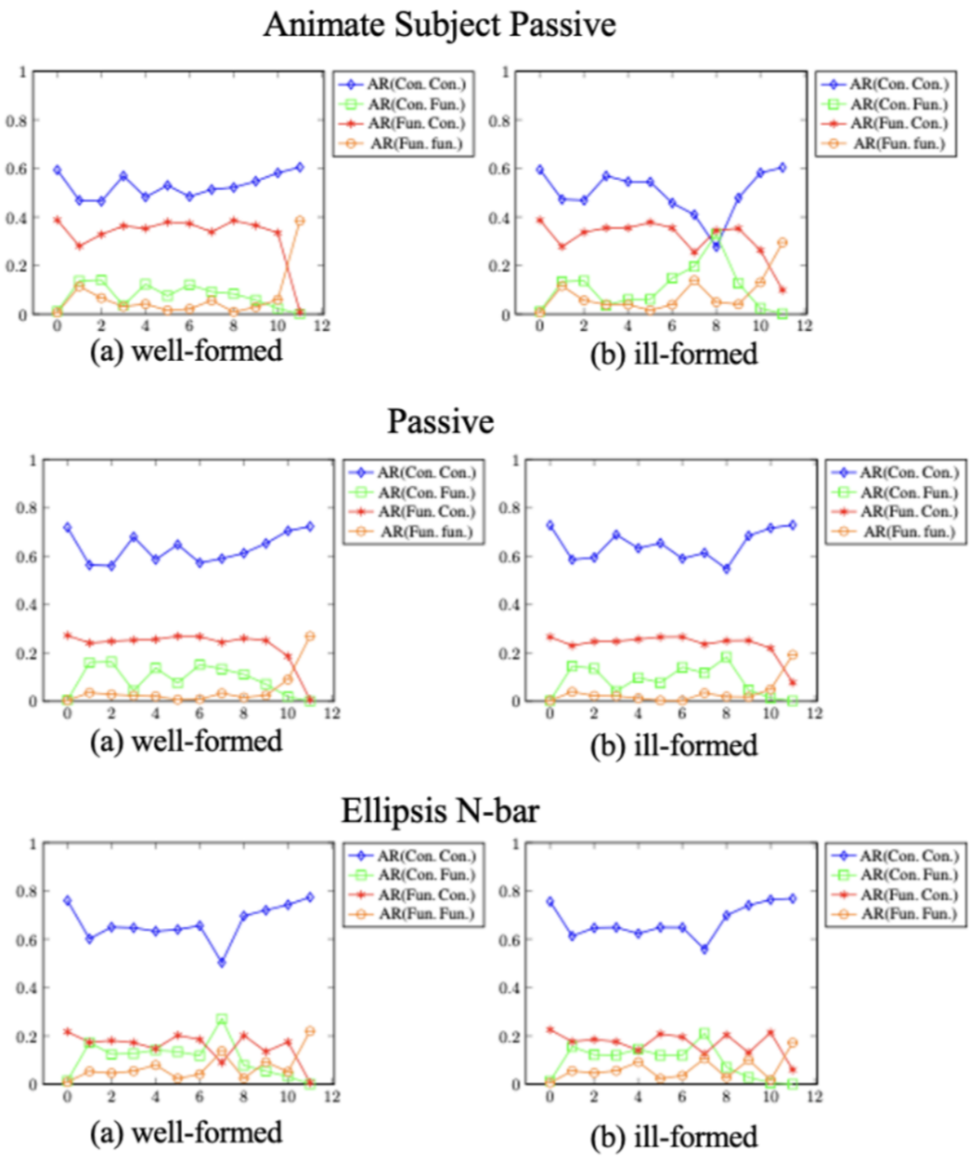
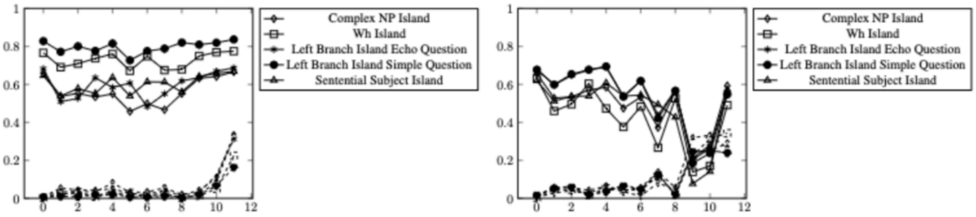


Figure 6.2: Lexical Category-based Affinity Relationship in Language Phenomena corresponding to Group I

6.2.2 Group II: Island Effects



(a) well-formed sentences (b) ill-formed sentences

Figure 6.3: The Affinity Relationship *(AR)* in five language phenomena associated with Group II: Complex NP Island, Wh Island, Left Branch Island Echo Question, Left Branch Island Simple Question, Sentential Subject Island. The solid lines in the figure represent the *AR(Con., Con.)*, while the dotted lines depict the *AR(Fun., Fun.)*.

The Affinity Prober successfully extracts distinct patterns specific to the Island Effect. Group II comprises five language phenomena: Complex NP Island, Wh Island, Left Branch Island Echo Question, Left Branch Island Simple Question, and Sentential Subject Island. Notably, *AR(Con., Con.)* exhibits different patterns between well-formed sentences and ill-formed sentences. In well-formed sentences, the Affinity Ratio in *AR(Con., Con.)* remains stable compared to ill-formed sentences, maintaining consistently high relationships. Similar to Group I, the relationship patterns between lexical categories within well-formed sentences are quite similar. We observe a sharp decline from Layer 1 to Layer 2 in *AR(Con., Con.)*, indicating a stronger focus on function words in the middle layers as the relationships between content words weaken. Analyzing *AR(Fun., Fun.)*, we find that the relationship between function words strengthens towards the final layer, indicating an effort to capture syntactic information. The patterns in *AR(Con., Con.)* for ill-formed sentences in Group II are particularly interesting. Layer 10 and 11 show a significant decline, with a strong emphasis on function words. Conversely, attention towards function words rapidly increases in the same layers, resulting in a cross pattern between the two graphs. Interpreting this in line with section 3.4.2, we understand that the language model struggles with structures differing from those observed and indirectly learned during forward propagation of ill-formed sentences. The sudden decline in layers 10 and 11 reflects this behavior, which can be attributed to the violation of NP island constraints and the complete disruption of sentence structure often seen in the Island Effect language phenomena.

When analyzing the linguistic phenomena of Complex NP Island, Wh Island, Left Branch Island Echo Question, Left Branch Island Simple Question, and Sentential Subject Island, there are some findings (Figure 6.4).

Complex NP Island Both well-formed and ill-formed sentences show fluctuations in the *AR(Con., Fun.)* patterns, starting low and decreasing towards the end. ill-formed sentences generally exhibit higher values and a broader range of fluctuations, suggesting a potentially stronger interaction. Specific layers, like layer 10, demonstrate distinct interactions in ill-formed sentences. The *AR(Fun., Con.)* patterns also exhibit fluctuations, with ill-formed sentences showing slightly higher values overall and specific layers of note.

Wh Island The strength and direction of *AR(Con., Fun.)* can vary between the relationships. ill-formed sentences tend to have higher overall values, indicating a stronger interaction. Significant differences exist in specific layers, such as layer 10, where ill-formed sentences display much higher values. In *AR(Fun., Con.)*, ill-formed sentences generally have higher values, suggesting a stronger association.

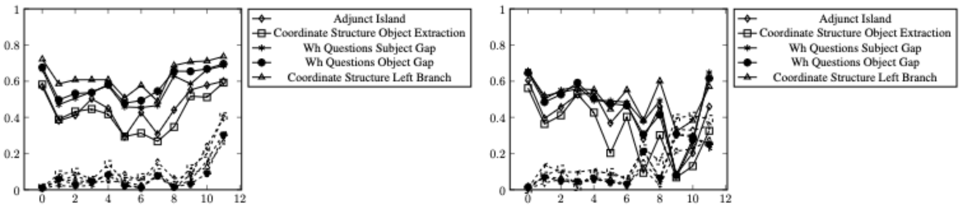
Left Branch Island Echo Question Both well-formed and ill-formed sentences exhibit higher values in *AR(Con., Fun.)* for specific layers, indicating a relatively stronger relationship. ill-formed sentences generally have higher overall values, especially in layer 8. In *AR(Fun., Con.)*, ill-formed sentences also tend to have higher values overall, with notable differences in specific layers.

Left Branch Island Simple Question There is a presence of positive values across all layers in *AR(Con., Fun.)* for both well-formed and ill-formed sentences. ill-formed sentences have higher overall values, particularly in layers 1 to 9, while well-formed sentences show stronger association in layers 10 to 12. In *AR(Fun., Con.)*, ill-formed sentences again have higher values overall, with variations in specific layers.

Sentential Subject Island There are differences in overall values and specific layers between the patterns. ill-formed sentences tend to have higher values in *AR(Con., Fun.),* particularly in layer 10. In *AR(Fun., Con.)*, there is no significant difference in overall values, but layer 10 shows higher values in well-formed sentences.

Figure 6.4: Lexical Category-based Affinity Relationship in Language Phenomena corresponding to Group II

6.2.3 Group III: Syntactic Constraints on Movement



(a) well-formed sentences (b) ill-formed sentences

Figure 6.5: The Affinity Relationship *(AR)* in five language phenomena associated with Group III: Adjunct Island, Coordinate Structure Object Extraction, Wh Questions Subject Gap, Wh Questions Object Gap, Coordinate Structure Left Branch. The solid lines in the figure represent the *AR(Con., Con.)*, while the dotted lines depict the *AR(Fun., Fun.)*.

Group III is composed of linguistic phenomena with constraints on movement. Among them, Adjunct Island is also included. The first difference from Group II is that the ratio of *AR(Con., Con)* relationships is lower in Group III compared to Group II. Secondly, there is activation of *AR(Fun., Fun.)* at the 8^{th} layer. According to section 3.4.2, the 8^{th} layer showed a tendency to give high attention to function words. Except for previous Ellipsis N-bar, in well-formed sentences, there is no significant increase in the reinforcement of function words at the 8^{th} layer. In contrast, Group III exhibits patterns that most closely match the tendencies observed in the layer analysis of section 3.4.2. The *AR(Con., Con.)* relationship undergoes a sharp drop from the 1st layer to the 2^{nd} layer, while the *AR(Fun., Fun.)* relationship increases simultaneously. As mentioned earlier, there is a drastic rise in function words at the 8^{th} layer, indicating a decrease in attention to content words. Near the last layer, which shows a high tendency towards content words, the *AR(Con., Con.)* relationship remains stable. Similar to other groups, there is a significant rise in

AR(Fun., Fun.) near the last layer, and the gap difference between *AR(Con., Con.)* is not significant. In contrast, in ill-formed sentences, there is an increase in function words at the 8th layer, but the relationship between *AR(Con., Con.)* is highly irregular, and there is significant fluctuation. This phenomenon, similar to Group II, reflects the difficulty of BERT in correctly interpreting syntactic dependency when processing sentences that violate constraints in movement, making it challenging to focus on which lexical category. This is clearly demonstrated in AR(Con., Con.). Our result analysis is stronger and more reliable based on the findings of section 3.4.2.

When analyzing the linguistic phenomena of Adjunct Island, Coordinate Structure Object Extraction, Wh Questions Subject Gap, Wh Questions Object Gap, and Coordinate Structure Left Branch, several findings emerge (Figure 6.6).

Adjunct Island The *AR(Con., Fun.)* patterns show fluctuations and non-linear trends. ill-formed sentences have a larger range of AR values, with higher maximum values, compared to well-formed sentences. Additionally, there is a spike at layer 9 in ill-formed sentences. The final AR value at layer 12 is notably higher for ill-formed sentences compared to well-formed sentences. In *AR(Fun., Con.)*, both patterns exhibit fluctuating trends, but there are differences in the lowest values and the final *AR(Fun., Con.)* at layer 12, with ill-formed sentences having lower values at specific layers and at the end of the series. There is also a significant decline at layer 9 in ill-formed sentences.

Coordinate Structure Object Extraction Both well-formed and ill-formed sentences show decreasing values in *AR(Con., Fun.)* and AR(Fun., Con.) patterns from layer 1 to layer 12. ill-formed sentences tend to have slightly higher values, with distinct AR patterns at certain layers. The range of fluctuations is similar, but ill-formed sentences exhibit higher peaks and more pronounced decreases compared to well-formed sentences.

Wh Questions Subject Gap The *AR(Con., Fun.)* patterns show higher overall values in ill-formed sentences compared to well-formed sentences. Specific layers, like layer 10, exhibit significantly higher values in ill-formed sentences, indicating a stronger *AR(Con., Fun.)* in ill-formed sentences at those points. In *AR(Fun., Con.)*, while the overall values are relatively similar, there are differences in specific layers, such as layer 2, where well-formed sentences have significantly higher values. This suggests a stronger *AR(Fun., Con.)* in well-formed sentences, particularly in layer 2.

Wh Questions Object Gap Ill-formed sentences have higher overall values in *AR(Con., Fun.)* compared to well-formed sentences, indicating a stronger relationship in ill-formed sentences. Significant differences are observed in specific layers, such as layer 10, where ill-formed sentences have considerably higher values. In *AR(Fun., Con.)*, ill-formed sentences also tend to have slightly higher overall values, with slight variations in specific layers, such as layer 8. This suggests a weaker *AR(Fun., Con.)* in ill-formed sentences for that specific layer.

Coordinate Structure Left Branch Both well-formed and ill-formed sentences exhibit fluctuating values in the *AR(Con., Fun.)* patterns. ill-formed sentences generally have higher values, particularly in certain layers like layer 6 and 8. The overall trend in *AR(Con., Fun.)* for ill-formed sentences shows higher peaks and more pronounced variations compared to well-formed sentences. In *AR(Fun., Con.)*, ill-formed sentences also tend to have higher values, with distinct patterns in specific layers, indicating a potentially stronger relationship between *AR(Con., Fun.)* in ill-formed sentences.

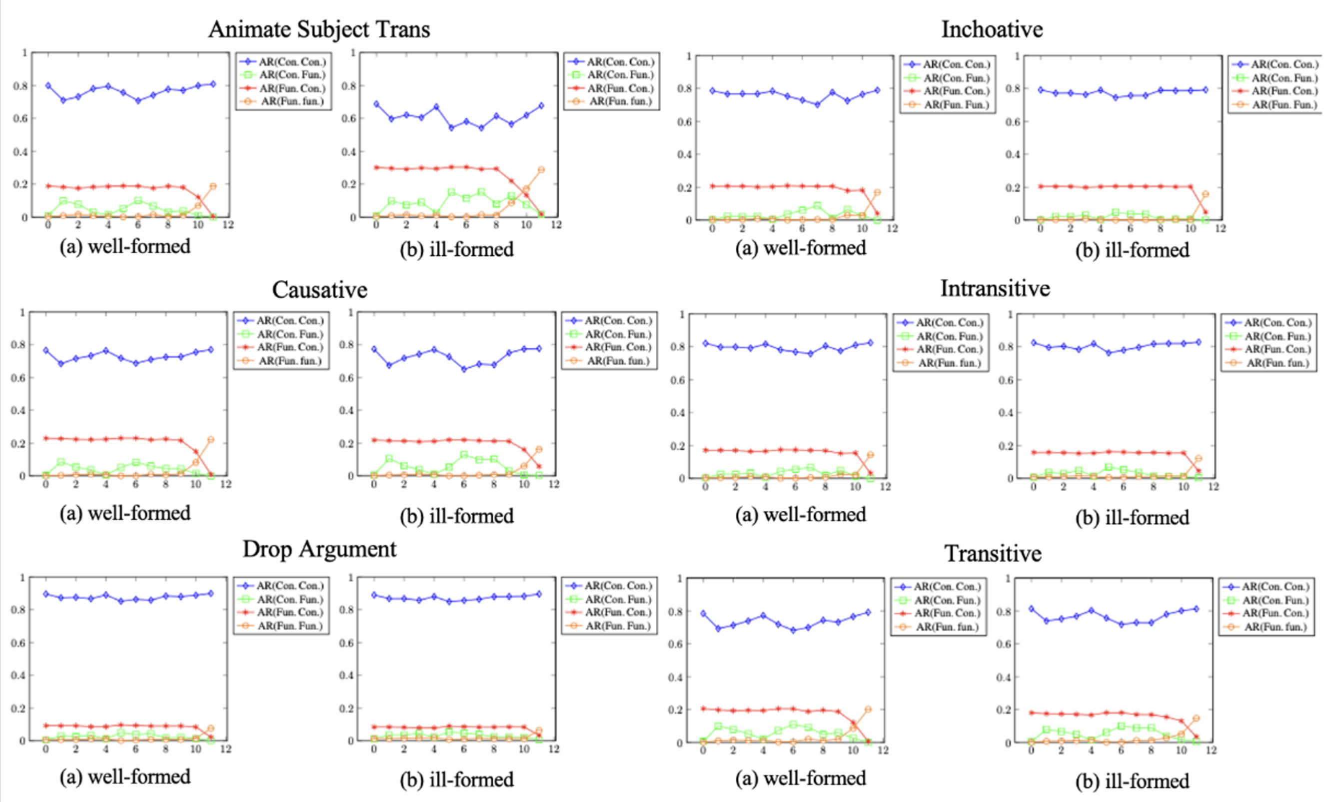
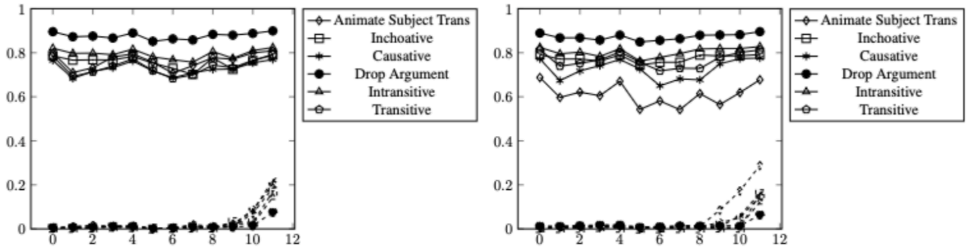


Figure 6.6: Lexical Category-based Affinity Relationship in Language Phenomena corresponding to Group III

6.2.4 Group IV: Verbal Predicate Types and Argument Structure



(a) well-formed sentences (b) ill-formed sentences

Figure 6.7: The Affinity Relationship *(AR)* in six language phenomena associated with Group III: Animate Subject Trans, Inchoative, Causative, Drop Argument, Intransitive, Transitive. The solid lines in the figure represent the *AR(Con., Con.)*, while the dotted lines depict the *AR(Fun., Fun.)*.

Group IV typically focuses on different types of verbal predicates and their argument structures. Their patterns exhibit a very consistent form, unlike other groups. The reason for this can be observed when examining example sentences of each linguistic phenomenon. In many cases, minimal pair sentences do not disrupt the sentence structure on the surface level. It is often in the lexical dimension where non-clauses are formed, with issues such as problems with the number of arguments taken by the predicate or semantically incorrect thematic roles. Therefore, BERT, in the process of learning information about minimal pairs through the CSS approach, compared sentences that do not have significant structural differences. Hence, the overall patterns of well-formed sentences and ill-formed sentences appear to be similar. However, there are also common patterns that emerge in Group IV. These include the initial downward trend of *AR(Con., Con)*, subtle reinforcement of the *AR(Fun., Fun.)* relationship at the 8th layer, and a rapid rise in the relationship between function words near the last layer. Similarly, there is a slight weakening of the relationship between content words at intermediate layers. A characteristic of Group IV is that there is not a significant difference in patterns between well-formed sentences and ill-formed sentences, unlike other groups. As mentioned above, Group IV discusses the grammaticality in the relationship between predicates and arguments, and the information about this is likely to be better represented in Word Embedding rather than Attention.

When examining the linguistic phenomena of Animate Subject Trans, Inchoative, and Causative, interesting patterns emerge in the *AR(Con., Fun.)* and *AR(Fun., Con.)* relationships between well-formed sentence and ill-formed sentence (Figure 6.8).

Animate Subject Trans The *AR(Con., Fun.)* patterns show similar variability but diverge in specific layers. ill-formed sentences generally have higher values from the 4^{th} layer, indicating a stronger *AR(Con., Fun.)* relationship. Significant differences are observed at the 12^{th} layer, where well-formed sentences drop to 0 while ill-formed sentences remain higher. Fluctuations at the 6^{th}, 8^{th}, and 10^{th} layers are seen in ill-formed sentences, not mirrored in well-formed sentences. In terms of *AR(Fun., Con.)*, both patterns start with high values and gradually decrease, but ill-formed sentences consistently have higher values across all layers, suggesting a stronger *AR(Fun., Con.)* relationship. The rate of decrease also varies, with ill-formed sentences showing a more notable decline, especially after the 10^{th} layer. Notably, at the 12^{th} layer, well-formed sentences have a significantly lower value compared to ill-formed sentences. The drop at the 10^{th} layer is also more significant in ill-formed sentences.

Inchoative Both well-formed and ill-formed sentences exhibit a weak *AR(Con.,* *Fun.)* with low values. well-formed sentences tend to have slightly higher values, suggesting a potentially stronger *AR(Con., Fun.)* in well-formed sentences. In terms of *AR(Fun., Con.)*, both patterns show a similar overall trend, but specific values may vary slightly, indicating potential differences based on sentence grammaticality.

Causative Both well-formed and ill-formed sentences display similar dynamics in the *AR(Con., Fun.)* patterns, but differences arise in overall values, specific points, ending values, and peak values. ill-formed sentences generally have slightly higher values, indicating a stronger *AR(Con., Fun.)* interaction. Notably, ill-formed sentences end with a non-zero value at layer 12, suggesting continued *AR(Con., Fun.)* interaction. The peak value for ill-formed sentences occurs later compared to well-formed sentences, indicating intensification of the *AR(Con., Fun.)* interaction in ill-formed sentences. In terms of *AR(Fun., Con.)*, both patterns show a similar trend but differ in overall value, drop-off point, intermediate fluctuations, and initial values. well-formed sentences generally have slightly higher values, suggesting a stronger *AR(Fun., Con.)* interaction. The drop towards the end is more dramatic for well-formed sentences.

Drop Argument Both grammatical and ill-formed sentences exhibit a weak AR(Con., Fun.) with low and consistent values across most layers. However, differences arise in the overall values and specific layers. ill-formed sentences tend to have slightly higher values, suggesting a potentially stronger *AR(Con., Fun.)* in ill-formed sentences. Specific layers, such as 2, 3, and 4, show lower values in well-formed sentences compared to ill-formed sentences, indicating a weaker *AR(Con., Fun.)* in well-formed sentences for these layers. In terms of *AR(Fun., Con.)*, both patterns show a moderate *AR(Fun., Con.)* with relatively close values across most layers. well-formed sentences generally have slightly higher values, suggesting a potentially stronger *AR(Fun., Con.)* in well-formed sentences. However, the differences are minor, indicating a similar overall trend in *the AR(Fun., Con.)* relationship between the two sentence types.

Intransitive Both grammatical and ill-formed sentences exhibit a positive *AR(Con., Fun.)* with relatively consistent positive values across most layers. However, there are differences in magnitude between the patterns. ill-formed sentences tend to have higher values, indicating a stronger *AR(Con., Fun.)* in ill-formed sentences. Specific layers show variations, with some layers having higher values in the ungrammatical pattern and others in the grammatical pattern. Notably, the last layer has a lower value in the grammatical pattern compared to the ungrammatical pattern, suggesting a weaker *AR(Con., Fun.)* in well-formed sentences for this specific layer. In terms of *AR(Fun., Con.)*, both patterns exhibit a positive *AR(Fun., Con.)* with relatively close values across most layers. well-formed sentences generally have slightly higher values, indicating a potentially stronger *AR(Fun., Con.)* in well-formed sentences. However, there are slight variations in specific layers, such as at layer 12, where the value in the *AR(Fun., Con.)* on well-formed sentences is lower than in the *AR(Fun., Con.)* on ill-formed sentences, implying a weaker *AR(Fun., Con.)* in well-formed sentences for this specific layer.

Transitive Both well-formed and ill-formed sentences show a moderate *AR(Con., Fun.)* with relatively higher values in specific layers compared to other layers. The overall values are similar in magnitude, but there are differences in specific layers. Certain layers have higher values in the grammatical pattern, indicating a relatively stronger *AR(Con., Fun.)* in well-formed sentences for those layers, while other layers have higher values in the ungrammatical pattern, suggesting a relatively stronger *AR(Con., Fun.)* in ill-formed sentences for those layers. Notably, the last layer has a lower value in the grammatical pattern compared to the ungrammatical pattern, indicating a weaker *AR(Con., Fun.)* in well-formed sentences for this specific layer. In terms of *AR(Fun., Con.)*, both patterns exhibit a general trend of decreasing values as the layer increases. The overall values are relatively similar, but there are differences in specific layers. Some layers have slightly higher values in the grammatical pattern, indicating a relatively stronger *AR(Fun., Con.)* in well-formed sentences for those layers, while other layers have slightly higher values in the ungrammatical pattern, suggesting a relatively stronger AR(Fun., Con.).

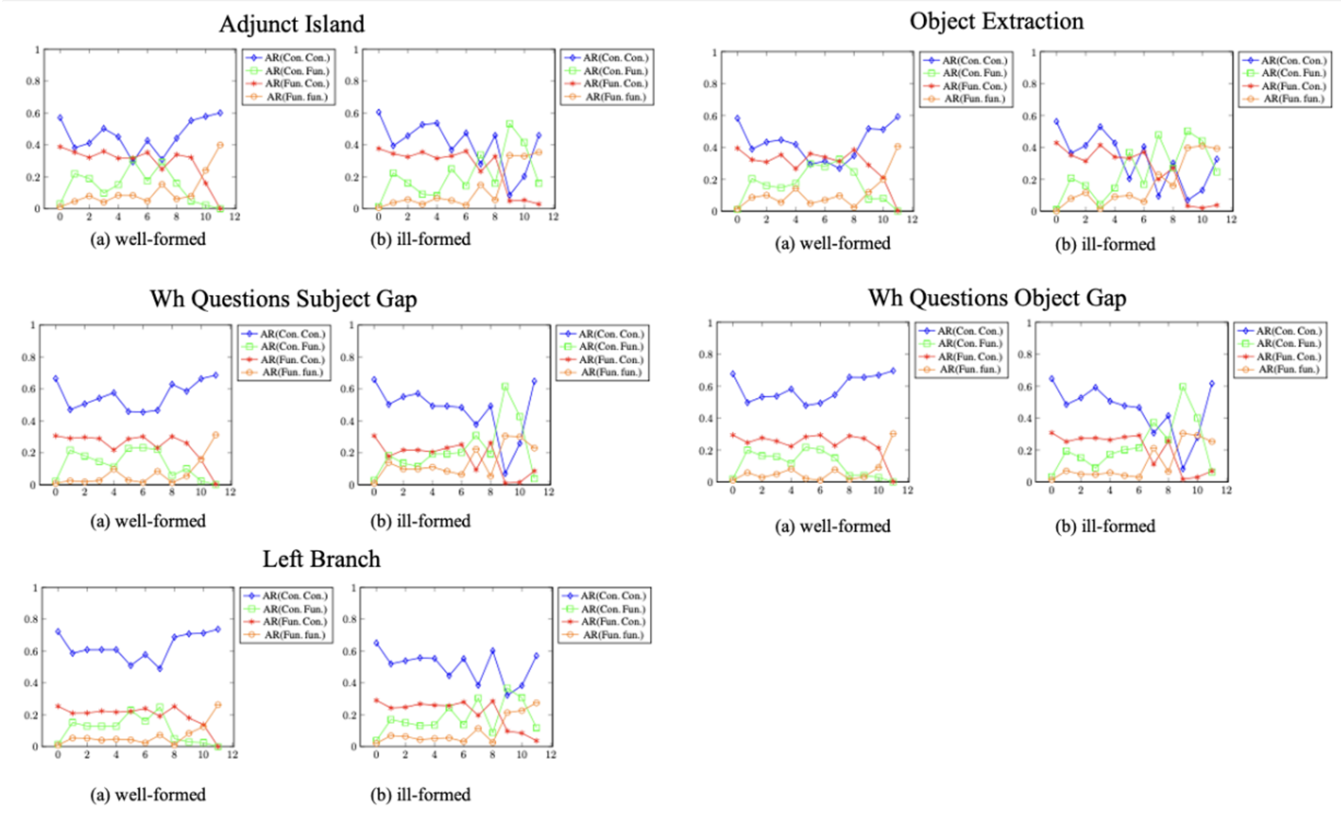


Figure 6.8: Lexical Category-based Affinity Relationship in Language Phenomena corresponding to Group IV

6.3 Summary

In this chapter, we apply the Affinity Prober to interpret patterns obtained from the Affinity Relationship, $AR(c_1, c_2)$, in each layer of the BERT model, considering different linguistic phenomena. It is found that not all linguistic phenomena have distinct patterns across layers; some phenomena show similar patterns that can be clustered together. Four distinct groups were identified, demonstrating the interaction between linguistic phenomena and layer-wise patterns in the *AR*, which helps deepen our understanding of language processing in BERT.

We observed common patterns in the layers of the BERT model across four groups. These patterns indicate a shift in attention from content words to function words during BERT's layer-wise processing. This finding aligns perfectly with the Layer tendency identified by the advanced ADTRAS algorithms. As a result, we can generalize that the First layer is content word-friendly, while the second layer is function word-friendly. Furthermore, we noticed a consistent trend in all groups where the attention ratio once again favors content words in the final layer. This observation corresponds well with the valuable insights provided by the ADTRAS algorithm, which identifies layers 10, 11, and 12 as content word-friendly layers. The middle layers of Groups 1, 2, and 3 particularly emphasize function words, suggesting a heightened focus on the functional aspects of sentences, such as grammar and syntactic relationships. This correlation supports the significance of layer 8 as a function word-friendly layer, as indicated by the output of the ADTRAS algorithm. Importantly, the final layers, known for their content word-friendly attributes, effectively address the interplay among function words. This intriguing phenomenon can be attributed to the indirect assimilation of the relevance of function words through fine-tuning on the CSS approach. The final layers are specialized for task-specific objectives, which contributes to their ability to rectify the role of function words.

We have successfully utilized the insightful patterns extracted from the Affinity Prober to cluster and explain various linguistic phenomena. These clusters exhibit fascinating interconnections, including the Island effect, movement constraint, Verb and Argument. Furthermore, we have established a strong link between the output of the ADTRAS algorithm and the layer tendencies discovered through the Affinity Prober, demonstrating the coherence and robustness of our analysis.

Chapter 7. Conclusion

In this study, we aimed to enhance the methodology proposed by Jang et al (2022) through additional experiments and analysis. We introduced the ADTRAS algorithm, which analyzes patterns at each layer of the BERT model and improves the interpretability of token relationships within attention scores. Through empirical experiments, we provided evidence that BERT autonomously learns linguistic knowledge related to lexical categories. We also investigated the general tendencies of BERT's layers when processing content words and function words, highlighting its processing characteristics associated with different word types.

Furthermore, we examined patterns in syntactic linguistic phenomena processed by BERT, focusing on specific phenomena within the BLiMP dataset. Our analysis revealed the potential of the Affinity Prober in understanding syntactic structures processed by BERT and facilitated clustering of similar linguistic phenomena. While this study offers valuable insights, it is important to acknowledge its limitations.

- First, our analysis focuses primarily on syntactic linguistic phenomena, neglecting other aspects of phenomena such as semantic, morphology, or discourses. Future research should aim to incorporate a broader range of linguistic phenomena to provide a more comprehensive understanding of BERT's capabilities.
- Second, our study relies on the use of the BERT model and the specific datasets employed, namely GLUE, SuperGLUE, and BLiMP. The findings may not necessarily generalize to other language models or datasets. Therefore, caution should be exercised when extrapolating the results beyond the scope of this study.
- Third, while the ADTRAS algorithm improves interpretability, it still relies on part-of-speech, which have inherent limitations in capturing complex linguistic relationships. Future research could explore alternative approaches or combine Affinity Prober with other linguistic features to gain deeper insights into BERT's processing mechanisms.
- Lastly, the Affinity Prober clusters linguistic phenomena based on patterns observed in BERT's layers. While this approach provides valuable information, it is important to note that clustering alone does not imply causal relationships or deeper understanding of linguistic phenomena. Further investigations and complementary analysis are needed to validate and interpret the observed patterns more thoroughly.

By addressing these limitations, researchers can further refine our understanding of BERT and its applications in natural language processing.

References

Jang, D., Kim, E., & Shin, H. Analysis on the elements of the decision boundaries of linguistic acceptability in Language Model using Affinity Prober. *Korean journal of Linguistics*. 47, 829-855 (2022)

Liu, N., Gardner, M., Belinkov, Y., Peters, M. & Smith, N. Linguistic knowledge and transferability of contextual representations. *ArXiv Preprint ArXiv:1903.08855.* (2019)

Hewitt, J. & Manning, C. A structural probe for finding syntax in word representations. Proceedings Of *The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers)*. pp. 4129-4138 (2019)

Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H. & Glass, J. What do neural machine translation models learn about morphology?. *ArXiv Preprint ArXiv:1704.03471*. (2017)

Conneau, A., Kruszewski, G., Lample, G., Barrault, L. & Baroni, M. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *ArXiv Preprint ArXiv:1805.01070.* (2018)

Hupkes, D., Veldhoen, S. & Zuidema, W. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal Of Artificial Intelligence Research. 61* pp. 907-926 (2018)

Hewitt, J. & Liang, P. Designing and interpreting probes with control tasks. *ArXiv Preprint ArXiv:1909.03368.* (2019)

Kim, T., Choi, J., Edmiston, D. & Lee, S. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. *ArXiv Preprint ArXiv:2002.00737*. (2020)

Shen, Y., Lin, Z., Huang, C. & Courville, A. Neural language modeling by jointly learning syntax and lexicon. *ArXiv Preprint ArXiv:1711.02013.* (2017)

Klein, D. & Manning, C. A generative constituent-context model for improved grammar induction. *Proceedings Of The 40th Annual Meeting Of The Association For Computational Linguistics.* pp. 128-135 (2002)

Klein, D. & Manning, C. Corpus-based induction of syntactic structure: Models of dependency and constituency. *Proceedings Of The 42nd Annual Meeting Of The Association For Computational Linguistics (ACL-04).* pp. 478-485 (2004)

Klein, D. & Manning, C. Natural language grammar induction using a constituent-context model. *Advances In Neural Information Processing Systems.* 14 (2001)

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature.* 521, 436-444 (2015)

Kuncoro, A., Ballesteros, M., Kong, L., Dyer, C., Neubig, G. & Smith, N. What do recurrent neural network grammars learn about syntax?. *ArXiv Preprint ArXiv:1611.05774*. (2016)

Li, Z., Zhou, Q., Li, C., Xu, K. & Cao, Y. Improving BERT with syntax-aware local attention. *ArXiv Preprint ArXiv:2012.15150.* (2020)

Bai, J., Wang, Y., Chen, Y., Yang, Y., Bai, J., Yu, J. & Tong, Y. Syntax-BERT: Improving pre-trained transformers with syntax trees. *ArXiv Preprint ArXiv:2103.04350.* (2021)

Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. ArXiv Preprint ArXiv:1409.0473. (2014)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems.* 30 (2017)

Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805.* (2018), https://arxiv.org/pdf/1810.04805.pdf

Carpenter, P., Just, M., Rayner, K. & Raynor, K. Eye movements in reading: Perceptual and language processes. *What Your Eyes Do While Your Mind Is Reading*. pp. 275-305 (1983)

Neville, H. & Debra, L. Mills, and Donald S. Lawson. 1992. Fractionating language: Different neural subsystems with different sensitive periods. *CerebralCortex. 2*, https://pubmed.ncbi.nlm.nih.gov/1511223/

Jawahar, G., Sagot, B. & Seddah, D. What does BERT learn about the structure of language?. *ACL 2019-57th Annual Meeting Of The Association For Computational Linguistics*. (2019), https://aclanthology.org/P19-1356.pdf

Htut, P., Phang, J., Bordia, S. & Bowman, S. Do attention heads in BERT track syntactic dependencies?. *ArXiv Preprint ArXiv:1911.12246*. (2019),

Kovaleva, O., Romanov, A., Rogers, A. & Rumshisky, A. Revealing the dark secrets of BERT. *ArXiv Preprint ArXiv:1908.08593.* (2019), https://arxiv.org/pdf/1908.08593.pdf

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *ArXiv Preprint ArXiv:1804.07461.* (2018), https://arxiv.org/pdf/1804.07461.pdf

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances In Neural Information Processing Systems*. 32 (2019), https://arxiv.org/pdf/1905.00537.pdf

Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Transactions Of The Association For Computational Linguistics*. 8 pp. 842-866 (2021), https://arxiv.org/pdf/2002.12327.pdf

Liu, N., Gardner, M., Belinkov, Y., Peters, M. & Smith, N. Linguistic knowledge and transferability of contextual representations. *ArXiv Preprint ArXiv:1903.08855*. (2019), https://arxiv.org/pdf/1903.08855.pdf

Hao, Y., Dong, L., Wei, F. & Xu, K. Visualizing and understanding the effectiveness of BERT. *ArXiv Preprint ArXiv:1908.05620*. (2019), https://arxiv.org/pdf/1908.05620

Appendix

Table 8.1: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Adjunct Island Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NNP, NNP)$	8744
$AR(NN, NN)$	6775
$AR(IN, VBG)$	6135
$AR(VBG, VBG)$	5280
$AR(VBG, IN)$	4593
$AR(IN, IN)$	3714
$AR(DT, NN)$	3490
$AR(VBG, NNP)$	2099
$AR(WP, NN)$	1852
$AR(IN, NN)$	1802
$AR(WP, NNP)$	1646
$AR(WP, VBD)$	1613
$AR(NNS, NNS)$	1596
$AR(WP, WP)$	1425
$AR(VBG, NN)$	1349
$AR(DT, DT)$	1247
$AR(VBD, VBD)$	1231
$AR(IN, VBN)$	1189
$AR(VBN, VBN)$	1167
$AR(WP, VBZ)$	1120

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	12423
$AR(IN, NN)$	8910
$AR(WP, WP)$	3226
$AR(WP, NN)$	2986
$AR(NNP, NNP)$	2963
$AR(VBD, WP)$	2951
$AR(NNP, WP)$	2700
$AR(DT, NN)$	2294
$AR(IN, IN)$	2061
$AR(NN, WP)$	1856
$AR(VBZ, WP)$	1808
$AR(WP, VBD)$	1733
$AR(DT, WP)$	1575
$AR(VBD, NN)$	1552
$AR(VBG, NN)$	1552
$AR(MD, WP)$	1475
$AR(VBG, VBG)$	1409
$AR(NNS, NNS)$	1352
$AR(IN, WP)$	1339
$AR(NNP, VBG)$	1308

Table 8.2: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Animate Subject Passive Phenomenon

(a) well-formed

Affinity Relationship	Count
$AR(NN, NN)$	16533
$AR(DT, NN)$	13006
$AR(IN, NN)$	6255
$AR(IN, VBN)$	4130
$AR(VBN, NN)$	3056
$AR(VBN, IN)$	2910
$AR(NNP, NN)$	2510
$AR(DT, DT)$	1952
$AR(VBN, VBN)$	1890
$AR(IN, IN)$	1813
$AR(VBD, VBN)$	1666
$AR(NNS, NNS)$	1259
$AR(NN, DT)$	1234
$AR(NNP, NNP)$	1202
$AR(VBZ, VBN)$	1140
$AR(VBP, VBP)$	1096
$AR(VBD, NN)$	1053
$AR(NN, VBN)$	911
$AR(IN, DT)$	885
$AR(DT, NNS)$	716

(b) ill-formed

Affinity Relationship	Count
$AR(NN, NN)$	16640
$AR(DT, NN)$	12600
$AR(IN, NN)$	6631
$AR(VBN, IN)$	4131
$AR(IN, VBN)$	3697
$AR(VBN, NN)$	3044
$AR(NNP, NN)$	2499
$AR(IN, IN)$	1999
$AR(DT, DT)$	1737
$AR(VBD, VBN)$	1566
$AR(VBN, VBN)$	1504
$AR(DT, IN)$	1347
$AR(NNS, NNS)$	1187
$AR(NNP, NNP)$	1121
$AR(VBD, NN)$	1095
$AR(VBZ, VBN)$	1031
$AR(IN, DT)$	1000
$AR(NN, DT)$	959
$AR(VBP, VBP)$	955
$AR(NN, IN)$	839

Table 8.3: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Animate Subject Trans Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	8428
$AR(NNP, NNP)$	8079
$AR(DT, NN)$	3729
$AR(NNP, NN)$	2446
$AR(NNS, NNS)$	1483
$AR(VBD, NN)$	1341
$AR(NNP, VBD)$	1189
$AR(NNP, VBZ)$	1182
$AR(VBD, NNP)$	1167
$AR(VBD, VBD)$	1070
$AR(IN, NN)$	1001
$AR(VBZ, VBZ)$	961
$AR(DT, DT)$	836
$AR(VBG, NN)$	815
$AR(DT, NNS)$	765
$AR(VB, VB)$	682
$AR(VBZ, NNP)$	620
$AR(VBZ, NN)$	566
$AR(NNP, NNS)$	564
$AR(NN, NNP)$	552

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	12132
$AR(DT, NN)$	8178
$AR(NNP, NNP)$	4290
$AR(NN, DT)$	2784
$AR(DT, DT)$	2453
$AR(VBD, NN)$	1653
$AR(NNS, NNS)$	1547
$AR(NN, VBD)$	1151
$AR(DT, NNP)$	1149
$AR(DT, NNS)$	1036
$AR(VBD, VBD)$	1016
$AR(IN, NN)$	881
$AR(VBG, NN)$	814
$AR(NN, VBZ)$	799
$AR(VBZ, VBZ)$	751
$AR(VBZ, NN)$	732
$AR(VB, VB)$	706
$AR(DT, VBD)$	684
$AR(VBG, VBG)$	654
$AR(VBD, NNP)$	638

Table 8.4: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Causative Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	11341
$AR(DT, NN)$	5883
$AR(NNP, NNP)$	4320
$AR(NNP, NN)$	2811
$AR(NNS, NNS)$	2271
$AR(VBD, NN)$	1744
$AR(DT, DT)$	1378
$AR(DT, NNS)$	1364
$AR(VBD, VBD)$	1183
$AR(IN, NN)$	1131
$AR(NNP, VBD)$	1030
$AR(NNP, VBZ)$	995
$AR(VBZ, VBZ)$	959
$AR(VBP, VBP)$	947
$AR(VBZ, NN)$	811
$AR(VBG, NN)$	780
$AR(VBG, VBG)$	733
$AR(VBP, NN)$	620
$AR(NN, DT)$	576
$AR(NNS, VBP)$	545

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10184
$AR(DT, NN)$	5577
$AR(NNP, NNP)$	3736
$AR(NNP, NN)$	2445
$AR(NNS, NNS)$	2055
$AR(VBD, NN)$	1969
$AR(NNP, VBD)$	1392
$AR(DT, NNS)$	1250
$AR(IN, NN)$	1091
$AR(DT, DT)$	1073
$AR(VBD, VBD)$	995
$AR(DT, VBD)$	888
$AR(NNP, VBZ)$	884
$AR(NN, VBD)$	832
$AR(NN, DT)$	724
$AR(VBG, NN)$	701
$AR(VBD, NNP)$	631
$AR(VBP, VBP)$	628
$AR(VBZ, VBZ)$	605
$AR(VBZ, NN)$	594

Table 8.5: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Complex NP Island Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	13484
$AR(NNP, NNP)$	9925
$AR(VBP, VBP)$	6088
$AR(WP, NN)$	5656
$AR(NNS, NNS)$	4283
$AR(WP, VBP)$	3789
$AR(VBD, VBD)$	2965
$AR(DT, NN)$	2838
$AR(WP, VBD)$	2441
$AR(VBD, NN)$	2178
$AR(NNP, NN)$	2053
$AR(VBZ, VBZ)$	2035
$AR(NN, NNP)$	1835
$AR(WP, WP)$	1806
$AR(IN, NN)$	1768
$AR(NNS, NNP)$	1433
$AR(WP, VBZ)$	1350
$AR(VBD, WP)$	1295
$AR(DT, DT)$	1264
$AR(IN, NNS)$	1220

(b) ill-formed

Affinity Relationship	**Count**
$AR(NNP, NNP)$	8521
$AR(NN, NN)$	8253
$AR(VBP, VBP)$	4495
$AR(NNS, NNS)$	4409
$AR(WP, WP)$	3923
$AR(WP, VBP)$	3803
$AR(VB, VB)$	3371
$AR(VBD, WP)$	2787
$AR(WP, VBD)$	2682
$AR(VBD, VBD)$	2610
$AR(NN, WP)$	2188
$AR(VBP, WP)$	2107
$AR(WP, NN)$	2091
$AR(IN, NN)$	2059
$AR(DT, NN)$	1995
$AR(NNP, WP)$	1937
$AR(NNS, WP)$	1832
$AR(NN, NNP)$	1805
$AR(DT, WP)$	1731
$AR(WP, VB)$	1428

Table 8.6: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Coordinate Structure Left Branch Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	13256
$AR(NNP, NNP)$	6441
$AR(CC, NN)$	5385
$AR(NNP, NN)$	4072
$AR(NNS, NNS)$	4008
$AR(VBZ, VBZ)$	2588
$AR(VBP, VBP)$	2309
$AR(CC, CC)$	2084
$AR(JJ, JJ)$	2075
$AR(WP, NN)$	1935
$AR(DT, NN)$	1926
$AR(VBD, NN)$	1865
$AR(JJ, NN)$	1860
$AR(NNP, CC)$	1697
$AR(NNS, JJ)$	1463
$AR(JJ, NNS)$	1430
$AR(VBD, VBD)$	1355
$AR(NN, CC)$	1321
$AR(VBZ, NN)$	1310
$AR(CC, VBP)$	1074

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10912
$AR(NNP, NNP)$	6419
$AR(CC, NN)$	4901
$AR(NNP, NN)$	3543
$AR(NNS, NNS)$	3430
$AR(NNS, CC)$	2901
$AR(VBP, VBP)$	1996
$AR(DT, NN)$	1879
$AR(NNP, CC)$	1784
$AR(VBD, WDT)$	1452
$AR(CC, CC)$	1443
$AR(VBD, NN)$	1400
$AR(CC, NNS)$	1347
$AR(WDT, WDT)$	1309
$AR(NNS, NN)$	1250
$AR(WP, WP)$	1225
$AR(NNP, VBD)$	1111
$AR(CC, NNP)$	1088
$AR(NNS, VBG)$	1059
$AR(VBG, VBG)$	1051

Table 8.7: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Coordinate Structure Object Extraction Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10652
$AR(NNP, CC)$	7422
$AR(WP, NN)$	6554
$AR(NNP, NNP)$	5647
$AR(CC, NN)$	4064
$AR(VBP, VBP)$	3039
$AR(CC, CC)$	2891
$AR(VBD, NN)$	2739
$AR(VBP, NN)$	2544
$AR(WP, VBP)$	2312
$AR(NNP, NN)$	2268
$AR(CC, NNP)$	2255
$AR(DT, NN)$	1894
$AR(MD, NN)$	1887
$AR(NNS, NNS)$	1164
$AR(DT, CC)$	1083
$AR(NNS, CC)$	1074
$AR(WP, WP)$	1053
$AR(NN, CC)$	973
$AR(CC, VBP)$	887

(b) ill-formed

Affinity Relationship	**Count**
$AR(NNP, NNP)$	7219
$AR(CC, NNP)$	4691
$AR(NN, NN)$	3808
$AR(WP, WP)$	2943
$AR(NNP, WP)$	2821
$AR(CC, WP)$	2696
$AR(VBD, WP)$	2259
$AR(DT, NN)$	2093
$AR(NN, CC)$	1955
$AR(WP, NNP)$	1880
$AR(VBZ, WP)$	1718
$AR(NNP, CC)$	1706
$AR(CC, NN)$	1617
$AR(WP, VBD)$	1496
$AR(MD, WP)$	1433
$AR(WP, VBZ)$	1372
$AR(VBN, CC)$	1366
$AR(NN, WP)$	1209
$AR(WP, MD)$	1118
$AR(DT, WP)$	1111

Table 8.8: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Drop Argument Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10232
$AR(NNP, NN)$	3860
$AR(NNP, NNP)$	2766
$AR(VBP, VBP)$	1738
$AR(VB, VB)$	1512
$AR(NNS, NNS)$	1489
$AR(JJ, JJ)$	1314
$AR(DT, NN)$	1248
$AR(NNS, VBP)$	975
$AR(MD, VB)$	677
$AR(VBD, NN)$	660
$AR(DT, VBP)$	507
$AR(NN, NNP)$	453
$AR(IN, NN)$	438
$AR(VBN, VBN)$	435
$AR(NNP, VB)$	363
$AR(VBD, JJ)$	335
$AR(NNS, NNP)$	327
$AR(VBZ, VBZ)$	319
$AR(NNP, JJ)$	314

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	11699
$AR(NNP, NN)$	3686
$AR(NNP, NNP)$	2798
$AR(VBD, NN)$	2217
$AR(NNS, NNS)$	1858
$AR(DT, NN)$	1589
$AR(VBP, VBP)$	1320
$AR(VB, VB)$	1045
$AR(JJ, JJ)$	853
$AR(NNS, VBP)$	828
$AR(NNS, NN)$	736
$AR(VBG, NN)$	715
$AR(VBZ, NN)$	561
$AR(VBP, NN)$	555
$AR(VB, NN)$	510
$AR(IN, NN)$	498
$AR(NN, NNP)$	470
$AR(NNP, NNS)$	435
$AR(RP, RP)$	418
$AR(MD, VB)$	406

Table 8.9: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Ellipsis N-bar Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	17969
$AR(NNP, NNP)$	9348
$AR(NNS, NNS)$	6782
$AR(DT, NN)$	4701
$AR(JJ, JJ)$	4336
$AR(CC, NN)$	3750
$AR(IN, NN)$	3685
$AR(JJ, NNS)$	3262
$AR(JJ, NN)$	2884
$AR(CC, CC)$	2759
$AR(VBP, VBP)$	2594
$AR(VBD, NN)$	2227
$AR(CD, JJ)$	1952
$AR(IN, IN)$	1913
$AR(NNS, JJ)$	1829
$AR(NN, JJ)$	1807
$AR(NNP, CC)$	1802
$AR(VBD, VBD)$	1772
$AR(CC, NNP)$	1766
$AR(DT, DT)$	1744

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	16859
$AR(NNS, NNS)$	9352
$AR(NNP, NNP)$	8451
$AR(DT, NN)$	5556
$AR(CD, NN)$	4639
$AR(JJ, NN)$	3506
$AR(CC, NN)$	3315
$AR(IN, NN)$	3197
$AR(JJ, NNS)$	3053
$AR(CC, CC)$	2563
$AR(CD, NNS)$	2369
$AR(VBD, NN)$	2358
$AR(IN, NNS)$	2202
$AR(JJ, JJ)$	2050
$AR(VBP, VBP)$	1984
$AR(CC, NNS)$	1945
$AR(NNP, NN)$	1906
$AR(NNS, NN)$	1890
$AR(CC, NNP)$	1842
$AR(DT, NNS)$	1814

Table 8.10: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Inchoative Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10643
$AR(DT, NN)$	3721
$AR(VBP, VBP)$	2796
$AR(NNS, NNS)$	2445
$AR(VB, VB)$	1758
$AR(NNS, VBP)$	1589
$AR(NNP, NNP)$	1049
$AR(NNP, NN)$	1046
$AR(JJ, JJ)$	1021
$AR(DT, VBP)$	1009
$AR(DT, NNS)$	842
$AR(DT, DT)$	696
$AR(DT, VB)$	624
$AR(MD, VB)$	590
$AR(NN, DT)$	575
$AR(VBD, NN)$	545
$AR(NNS, VB)$	518
$AR(NNS, NN)$	460
$AR(DT, JJ)$	456
$AR(VBN, VBN)$	374

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	11052
$AR(DT, NN)$	3877
$AR(VBP, VBP)$	2497
$AR(NNS, NNS)$	2200
$AR(VB, VB)$	1802
$AR(NNS, VBP)$	1616
$AR(NNP, NN)$	1140
$AR(JJ, JJ)$	1073
$AR(DT, VBP)$	995
$AR(VBD, NN)$	989
$AR(NNP, NNP)$	940
$AR(DT, NNS)$	751
$AR(DT, VB)$	684
$AR(NNS, NN)$	627
$AR(NNS, VB)$	613
$AR(MD, VB)$	559
$AR(DT, DT)$	530
$AR(DT, JJ)$	526
$AR(VBP, VB)$	422
$AR(VBP, NN)$	395

Table 8.11: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Intransitive Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	9261
$AR(DT, NN)$	2683
$AR(VBP, VBP)$	2661
$AR(NNS, NNS)$	2333
$AR(NNP, NNP)$	2005
$AR(NNP, NN)$	1908
$AR(VB, VB)$	1819
$AR(NNS, VBP)$	1545
$AR(JJ, JJ)$	1449
$AR(DT, VBP)$	946
$AR(MD, VB)$	782
$AR(DT, NNS)$	712
$AR(VBD, NN)$	656
$AR(DT, DT)$	570
$AR(NNS, NN)$	518
$AR(DT, VB)$	511
$AR(NNS, VB)$	447
$AR(IN, NNS)$	445
$AR(NN, DT)$	428
$AR(VBN, VBN)$	397

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10890
$AR(DT, NN)$	3056
$AR(NNS, NNS)$	2206
$AR(NNP, NN)$	2061
$AR(NNP, NNP)$	1929
$AR(VBP, VBP)$	1855
$AR(VBD, NN)$	1709
$AR(VB, VB)$	1383
$AR(JJ, JJ)$	1348
$AR(NNS, VBP)$	1229
$AR(NNS, NN)$	992
$AR(VBG, NN)$	918
$AR(VBP, NN)$	811
$AR(DT, VBP)$	664
$AR(DT, NNS)$	609
$AR(MD, VB)$	584
$AR(DT, JJ)$	547
$AR(DT, VB)$	498
$AR(JJ, NN)$	457
$AR(VBZ, NN)$	444

Table 8.12: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Transitive Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	10184
$AR(NNP, NNP)$	9080
$AR(DT, NN)$	4342
$AR(NNS, NNS)$	2658
$AR(NNP, NN)$	1880
$AR(IN, NN)$	1415
$AR(DT, DT)$	1200
$AR(VBD, VBD)$	1156
$AR(VBD, NN)$	1119
$AR(VBP, VBP)$	1070
$AR(DT, NNS)$	1065
$AR(VBD, NNP)$	916
$AR(NNP, VBD)$	888
$AR(IN, NNP)$	888
$AR(VBZ, VBZ)$	845
$AR(NN, NNP)$	844
$AR(NN, DT)$	804
$AR(NNS, NNP)$	802
$AR(JJ, JJ)$	784
$AR(IN, NNS)$	775

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	9023
$AR(NNP, NNP)$	8160
$AR(DT, NN)$	3818
$AR(NNS, NNS)$	2330
$AR(NNP, NN)$	1673
$AR(NNP, VBD)$	1400
$AR(VBD, VBD)$	1309
$AR(VBD, NNP)$	1171
$AR(VBD, NN)$	1059
$AR(IN, NN)$	1029
$AR(NN, VBD)$	1013
$AR(DT, DT)$	976
$AR(DT, NNS)$	915
$AR(VBP, VBP)$	871
$AR(NN, DT)$	825
$AR(NNS, NNP)$	797
$AR(NN, NNP)$	749
$AR(DT, VBD)$	732
$AR(VBZ, VBZ)$	721
$AR(NNP, VBZ)$	696

Table 8.13: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Left Branch Island Echo Question Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	12664
$AR(WP, NN)$	3837
$AR(WDT, NN)$	3581
$AR(NNP, NNP)$	2337
$AR(NNP, NN)$	2129
$AR(WP, NN)$	2016
$AR(DT, NN)$	1546
$AR(VB, VB)$	1490
$AR(VBD, NN)$	1362
$AR(VBN, NN)$	1336
$AR(MD, NN)$	1272
$AR(VB, NN)$	1216
$AR(MD, VB)$	1198
$AR(VBG, VBG)$	1148
$AR(VBZ, NN)$	1113
$AR(NNP, VBZ)$	1106
$AR(NNP, VBD)$	971
$AR(VBG, NN)$	896
$AR(VBD, VBN)$	861
$AR(VBN, VBN)$	806

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	9977
$AR(NNP, NNP)$	3630
$AR(VBD, NN)$	2227
$AR(VBG, NN)$	1447
$AR(VB, NN)$	1332
$AR(IN, NN)$	1314
$AR(NN, WDT)$	1280
$AR(WP, WP)$	1265
$AR(WDT, WDT)$	1250
$AR(WP, NN)$	1243
$AR(VBD, WDT)$	1219
$AR(WDT, NN)$	1170
$AR(VB, VB)$	1151
$AR(NNP, NN)$	1101
$AR(NN, WP)$	1057
$AR(VBD, WP)$	999
$AR(MD, NN)$	955
$AR(VBZ, NN)$	954
$AR(NNP, VBD)$	917
$AR(DT, NN)$	874

Table 8.14: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Left Branch Island Simple Question Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	13375
$AR(NNP, NN)$	3672
$AR(NNP, NNP)$	3288
$AR(VBD, NN)$	3088
$AR(WP, NN)$	2943
$AR(VBZ, NN)$	2626
$AR(JJ, NN)$	2449
$AR(NNS, NNS)$	2126
$AR(VB, VB)$	1714
$AR(JJ, JJ)$	1569
$AR(VBZ, VBZ)$	1486
$AR(VBP, VBP)$	1468
$AR(DT, NN)$	1080
$AR(MD, NN)$	1015
$AR(NNS, NN)$	999
$AR(NNS, VBP)$	842
$AR(VBD, VBD)$	802
$AR(MD, VB)$	796
$AR(VB, NN)$	793
$AR(NNP, VB)$	785

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	9762
$AR(NNP, NNP)$	3311
$AR(VBD, NN)$	2534
$AR(VBG, NN)$	1534
$AR(WDT, WDT)$	1377
$AR(NN, WDT)$	1345
$AR(VBD, WDT)$	1330
$AR(WDT, NN)$	1243
$AR(IN, NN)$	1241
$AR(WP, WP)$	1228
$AR(WP, NN)$	1204
$AR(VB, NN)$	1102
$AR(NNP, NN)$	1077
$AR(VBD, WP)$	1063
$AR(VBZ, WDT)$	1026
$AR(NN, WP)$	1019
$AR(VBZ, NN)$	1015
$AR(NNP, VBD)$	977
$AR(VB, VB)$	921
$AR(VBN, NN)$	899

Table 8.15: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Passive Phenomenon

(a) well-formed

Affinity Relationship	Count
$AR(NNP, NNP)$	11996
$AR(NN, NN)$	8767
$AR(IN, NNP)$	4559
$AR(IN, VBN)$	4078
$AR(VBN, IN)$	3900
$AR(DT, NN)$	3423
$AR(NNS, NNS)$	3408
$AR(IN, NN)$	2760
$AR(VBN, NN)$	2218
$AR(NNP, NN)$	2105
$AR(VBN, VBN)$	2080
$AR(IN, IN)$	2062
$AR(VBD, VBN)$	1623
$AR(VBP, VBP)$	1490
$AR(DT, NNS)$	1323
$AR(IN, NNS)$	1139
$AR(DT, DT)$	1091
$AR(NN, NNP)$	1067
$AR(VBN, VBP)$	959
$AR(VBP, VBN)$	945

(b) ill-formed

Affinity Relationship	Count
$AR(NNP, NNP)$	10149
$AR(NN, NN)$	8553
$AR(IN, VBN)$	4688
$AR(IN, NNP)$	3620
$AR(VBN, IN)$	3154
$AR(DT, NN)$	3118
$AR(NNS, NNS)$	2743
$AR(IN, NN)$	2591
$AR(VBN, VBN)$	2549
$AR(NNP, NN)$	2419
$AR(VBN, NN)$	2158
$AR(VBD, VBN)$	2099
$AR(VBP, VBP)$	1434
$AR(NN, VBN)$	1351
$AR(VBP, VBN)$	1318
$AR(IN, IN)$	1144
$AR(NNP, IN)$	1125
$AR(DT, NNS)$	1118
$AR(VBZ, VBN)$	1097
$AR(IN, NNS)$	1011

Table 8.16: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Sentential Subject Island Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	15739
$AR(WP, NN)$	6288
$AR(DT, NN)$	4930
$AR(NNP, NN)$	3944
$AR(VBG, NN)$	3931
$AR(JJ, JJ)$	3632
$AR(VBG, VBG)$	2618
$AR(VBP, VBP)$	2492
$AR(NNS, NNS)$	2210
$AR(NNP, NNP)$	2040
$AR(WP, VBP)$	1841
$AR(VBD, NN)$	1836
$AR(DT, JJ)$	1764
$AR(DT, DT)$	1581
$AR(IN, NN)$	1551
$AR(VB, VB)$	1508
$AR(NNP, VBG)$	1377
$AR(DT, NNS)$	1362
$AR(VBG, JJ)$	1197
$AR(VBP, WP)$	1138

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	7816
$AR(NNP, NNP)$	4985
$AR(DT, NN)$	4350
$AR(VBG, VBG)$	3893
$AR(JJ, JJ)$	3406
$AR(VBD, WP)$	3060
$AR(WP, WP)$	2959
$AR(DT, WP)$	2729
$AR(VBG, NN)$	2572
$AR(NN, WP)$	2294
$AR(VBP, WP)$	2110
$AR(NNP, WP)$	1950
$AR(WP, VBD)$	1884
$AR(VBG, WP)$	1858
$AR(NNS, NNS)$	1821
$AR(MD, WP)$	1666
$AR(VBG, JJ)$	1646
$AR(DT, JJ)$	1605
$AR(IN, NN)$	1447
$AR(VB, VB)$	1436

Table 8.17: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Wh Island Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	5581
$AR(VB, VB)$	5407
$AR(VBD, VBD)$	5269
$AR(NNP, NNP)$	2555
$AR(WP, NN)$	2489
$AR(WP, VBD)$	2467
$AR(PRP, VB)$	2420
$AR(VBZ, VBZ)$	2373
$AR(WP, VB)$	2049
$AR(VBP, VBP)$	1976
$AR(PRP, VBD)$	1892
$AR(VBD, NN)$	1831
$AR(PRP, PRP)$	1723
$AR(NNP, VBD)$	1692
$AR(PRP, NN)$	1506
$AR(VBZ, NN)$	1363
$AR(MD, VB)$	1339
$AR(JJ, JJ)$	1267
$AR(WP, VBZ)$	1163
$AR(WP, WP)$	1123

(b) ill-formed

Affinity Relationship	**Count**
$AR(WP, WP)$	5037
$AR(NN, NN)$	4921
$AR(WP, VBP)$	4912
$AR(VBP, VBP)$	4801
$AR(WP, NN)$	3800
$AR(VBD, WP)$	3080
$AR(VB, VB)$	3042
$AR(WP, VBD)$	2984
$AR(NN, WP)$	2573
$AR(VBD, VBD)$	2319
$AR(WP, VB)$	2186
$AR(WP, VBZ)$	1962
$AR(VB, WP)$	1880
$AR(VBZ, WP)$	1855
$AR(VBG, WP)$	1835
$AR(NNP, WP)$	1823
$AR(VBP, WP)$	1799
$AR(NNP, NNP)$	1749
$AR(VBN, WP)$	1570
$AR(VBD, NN)$	1349

Table 8.18: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Wh Questions Object Gap Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	12620
$AR(DT, NN)$	5125
$AR(NNP, NNP)$	4957
$AR(NNS, NNS)$	4420
$AR(NNP, NN)$	3197
$AR(IN, NN)$	2857
$AR(VBP, VBP)$	2741
$AR(VB, VB)$	2596
$AR(DT, NNS)$	2571
$AR(DT, DT)$	2326
$AR(WDT, NN)$	2265
$AR(NN, DT)$	2047
$AR(VBD, VBD)$	1834
$AR(NNP, VBD)$	1635
$AR(VBD, NN)$	1609
$AR(VBZ, VBZ)$	1604
$AR(NNS, VBP)$	1582
$AR(DT, VBP)$	1433
$AR(NNP, VBZ)$	1389
$AR(IN, VBP)$	1354

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	12034
$AR(DT, NN)$	6799
$AR(NNP, NNP)$	4020
$AR(WP, WP)$	3712
$AR(DT, WP)$	3701
$AR(NN, WP)$	3698
$AR(NNP, WP)$	3443
$AR(NNS, NNS)$	3020
$AR(VBD, WP)$	2912
$AR(NNP, NN)$	2601
$AR(DT, NNS)$	1997
$AR(VBP, WP)$	1925
$AR(NNS, WP)$	1739
$AR(IN, WP)$	1712
$AR(VBD, NN)$	1682
$AR(NNP, VBD)$	1534
$AR(IN, NN)$	1501
$AR(VBZ, WP)$	1487
$AR(VB, WP)$	1359
$AR(DT, DT)$	1287

Table 8.19: Top 10 Frequency of Affinity Relationship Categorized by Part-of-Speech on Wh Questions Subject Gap Phenomenon

(a) well-formed

Affinity Relationship	**Count**
$AR(NNP, NNP)$	9576
$AR(NN, NN)$	8610
$AR(DT, NN)$	5452
$AR(NNS, NNS)$	4128
$AR(VBP, VBP)$	2569
$AR(DT, DT)$	2487
$AR(VBZ, VBZ)$	2464
$AR(DT, NNS)$	2363
$AR(VBD, VBD)$	2046
$AR(NN, DT)$	1851
$AR(WDT, VBD)$	1632
$AR(WDT, VBP)$	1585
$AR(WDT, WDT)$	1574
$AR(IN, NN)$	1416
$AR(VB, VB)$	1412
$AR(VBD, NNP)$	1392
$AR(WDT, VBZ)$	1325
$AR(NNS, WDT)$	1277
$AR(DT, VBD)$	1230
$AR(NNP, VBD)$	1224

(b) ill-formed

Affinity Relationship	**Count**
$AR(NN, NN)$	8245
$AR(NNP, NNP)$	7784
$AR(DT, WP)$	5760
$AR(DT, NN)$	4721
$AR(NN, WP)$	3748
$AR(WP, WP)$	3639
$AR(NNS, NNS)$	3340
$AR(VBD, WP)$	2609
$AR(NNS, WP)$	2344
$AR(DT, NNS)$	2067
$AR(WP, DT)$	2040
$AR(NNP, WP)$	1982
$AR(NN, DT)$	1880
$AR(VBP, WP)$	1875
$AR(VBP, VBP)$	1873
$AR(IN, WP)$	1704
$AR(NNP, NN)$	1330
$AR(VBN, WP)$	1295
$AR(VB, WP)$	1213
$AR(VBD, NN)$	1147

국문 초록

Transformer(Vaswani et al., 2017)의 등장 이후 Self Attention 기제를 사용한 다양한 사전학습 언어모델(Pre-trained Language Model)이 제안되었다. 이러한 사전학습 언어 모델은 일반적으로 미세조정(fine-tuning)을 통해 다양한 자연어 처리 문제에서 높은 성능을 보여왔다. 언어학 분야에서는 언어 모델의 내재적 언어 지식을 탐구하기 위해 통사론, 의미론, 언어 습득 등의 이론 및 실험 언어학 접근법을 기반으로 활발히 연구되고 있다. 본 논문은 Jang et al (2022)에서 제안한 언어 지식 탐침 방법론인 Affinity Prober의 사용 범주를 확장시키는 것을 목표로 한다. 이를 위해 self-attention mechanism에서 어텐션 스코어 값을 보존하며 토큰 간의 관계를 해석하는 알고리즘인 ADTRAS 알고리즘 (An Algorithm for Decrypting Token Relationships within Attention Scores)을 제안한다. 본 논문은 ADTRAS 알고리즘을 활용하여 첫 번째 실험에서 GLUE 벤치마크 내의 통사-의미적 기능을 요구하는 6가지 태스크에 각각 훈련된 BERT 모형의 레이어 패턴을 분석한다. 이를 통해 BERT 모형이 토큰 관계의 유의미한 변화를 포착하고, ADTRAS 알고리즘을 활용하여 BERT 어텐션 변화를 기반으로 BERT 모델이 스스로 어휘 범주(Lexical Category)를 활용하여 품사 정보를 학습한다는 실증적인 증거를 제시한다. 또한 어휘 범주를 중심으로 BERT 레이어의 분명한 언어학적 특징을 일반화한다. 두 번째 실험으로는 Affinity Prober를 활용하여 통사적 언어현상에서의 최소쌍 문장을 처리하는 BERT의 특징을 분석한다. 이 실험은 사용된 15가지의 통사적 언어현상이 BERT 모델에서 처리되는 과정을 Affinity Prober를 활용하여 탐구하여 레이어 별 패턴을 분석하는 것을 목적으로 한다. 이러한 실험 결과로 총 네 가지의 패턴이 관찰되었는데, 본 논문은 관찰된 패턴이 각각 유사한 언어현상 별로 묶인다고 주장한다. 첫 번째 패턴은 Passive와 Ellipsis N-bar와 관련된 언어현상들이 주를 이루며, 두 번째 패턴은 Island Effects, 세 번째 패턴은 Movement에서의 Syntactic Constraints에서의 언어현상, 마지막으로 네 번째 패턴에서는 Verb Predicate Types과 논항 구조에서의 언어현상들로 나타난다. 이러한 각 레이어 별 패턴이 ADTRAS 알고리즘에서의 결과와 일치한다는 점에서 본 실험을 통해 도출된 결과를 뒷받침한다. 요약하자면, 본 논문은 ADTRAS 알고리즘을

제안하고, Jang et al (2022)에서 제안한 Affinity Prober를 확장하여 연구에 활용하였다. 이 과정에서 통사적 언어현상의 BERT 레이어 별 패턴을 성공적으로 추출하여 결과를 설명하고자 노력하였다.