Linguistic Research 40(2): 317-352
DOI: 10.17250/khisli.40.2.202306.007

Grammatical illusions in BERT: Attraction effects of subject-verb agreement and reflexive-antecedent dependencies*

Ye-eun Cho
(Sungkyunkwan University)

Cho, Ye-eun. 2023. Grammatical illusions in BERT: Attraction effects of subject-verb agreement and reflexive-antecedent dependencies. *Linguistic Research* 40(2): 317-352. The phenomenon of attraction effects, whereby a verb erroneously retrieves a syntactically inaccessible but feature-matching noun, is a type of grammatical illusions (Phillips, Wagers, and Lau 2011) that can occur in long-distance subject-verb agreement in human sentence processing (Wagers et al. 2009). In contrast, reflexive-antecedent dependencies have been claimed to lack attraction effects when the reflexive and the antecedent mismatch (Dillon et al. 2013). Yet, some other studies have shown that attraction effects have been observed in reflexive-antecedent dependencies, when the number of feature mismatch between the reflexive and the antecedent increases (Parker and Philips 2017). These findings suggest that there are different cue weightings based on the predictability of the dependency, and these cues are combined according to different cue-combination scheme, such as a linear or a non-linear cue-combination rule (Parker 2019). These linguistic phenomena can be used to analyze how linguistic features are accessed and combined within the internal states of Deep Neural Network (DNN) language models. In the linguistic representations of BERT (Devlin et al. 2018), one of the pre-trained DNN language models, various types of linguistic information are encoded in each layer (Jawahar et al. 2019) and combined while passing through the layers. By measuring the performance of Masked Language Model (MLM), this study finds that both subject-verb agreement and reflexive-antecedent dependencies show attraction effects and follow the linear-combinatoric rule in BERT. The different results from human sentence processing suggest that the self-attention mechanism of BERT may not be able to capture the differences in the predictability of the dependency as effectively as memory retrieval mechanisms in humans. These findings have important implications for developing more understandable and interpretable explainable-AI (xAI) systems that better capture the complexities of human language processing. **(Sungkyunkwan University)**

* I thank two anonymous reviewers of *Linguistic Research* for their constructive feedback. I also thank Nayoun Kim, Seongmook Kim and Won Ik Cho for their very helpful discussion. This research was supported by the Sungkyunkwan University and the BK21 FOUR (Graduate School Innovation) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

Keywords xAI, DNN language models, BERT, Masked Language Model, cue-combinatoric scheme, attraction effects

1. Introduction

In recent years, artificial intelligence (AI) has made significant advances, resulting in increasingly complex and accurate algorithms and models. However, the complexity and opacity of these models have raised concerns about their trustworthiness and interpretability, leading to the emergence of explainable-AI (xAI) as a critical research area. While traditional xAI techniques focus on explaining the outputs of black-box models, they are often limited in their scope and do not provide a complete understanding of the underlying mechanisms of the model. Thus, there is a growing interest in developing future xAI systems that generate accurate, understandable and relevant explanations for human users (Pearl 2018; Holzinger et al. 2019), since it helps to build trust and confidence in the decisions made by AI. Accordingly, this study aims to improve the explainability and interpretability of deep neural network (DNN) language models by incorporating human sentence processing techniques.

In Natural Language Processing (NLP), most studies on assessing syntactic abilities of DNN language models have achieved high accuracy in prediction during long-distance dependency formation, including subject-verb agreement (Linzen et al. 2016; Gulordava et al. 2018; Marvin and Linzen 2018; Bacon and Regier 2019; Goldberg 2019; Lin, Tan, and Frank 2019) and reflexive-antecedent dependencies (Marvin and Linzen 2018; Goldberg 2019; Lin, Tan, and Frank 2019). However, the approaches to learn language of DNN language models are quite different from those of human. This cognitive difference of language models and humans can be a key factor to solve the black-box problem that deep learning poses (Perconti and Plebe 2020). Therefore, this study focuses not on the accuracy of dependency formation but on the way in which the model deals with linguistic information. By comparing how DNN language models process language to how humans process language, we can develop more transparent and interpretable xAI systems.

In human sentence processing, subject-verb agreement is one of the common examples of the dependencies. In subject-verb agreement, the verb must match the subject in number. Also, the noun which has the same morphological feature (e.g., number) as the verb must be at the syntactically licensed position (i.e., subject).

(1) a. **The key** to the cells **was** rusty.
    b. ***The key** to the cells **were** rusty.

In long-distance subject-verb agreement like (1) extracted from Wagers et al. (2009), although there is interference between the subject and the verb, the reader often achieves successful dependency formation, matching the verb (*was*) to the subject (*the key*) as in (1a). But sometimes the reader erroneously accepts ungrammatical sentences like (1b), linking the verb (*were*) to the syntactically incorrect target noun (*the cells*), since the number feature of the verb corresponds to that of the intervening noun, known as *attractor* (Bock and Miller 1991). This phenomenon of retrieving the grammatically illicit but feature-matching attractor in subject-verb dependency formation is regarded as *agreement attraction* (Pearlmutter, Garnsey, and Bock 1999; Wagers, Lau, and Phillips 2009; Dillon et al. 2013; Tanner, Nicol, and Brehm 2014; Lago et al. 2015; Tucker, Idrissi, and Almeida 2015; Kim, Brehm, and Yoshida 2019). Agreement attraction effects are referred to as a part of *grammatical illusions* (Phillips, Wagers, and Lau 2011), in which the reader erroneously accepts ungrammatical sentences for a brief time but realizes the ungrammaticality after careful consideration. The grammatical illusions in subject-verb agreement dependencies are claimed to occur as a result of using morphological features as retrieval cues in a content-addressable memory architecture, where linguistic information is indexed and accessed based on the feature content of the target element rather than its location.

In contrast, there have been different results shown in reflexive-antecedent dependencies. In reflexive-antecedent dependencies, the reflexive can be licensed when the features (e.g., number, gender, and animacy, etc.) of the antecedent and the reflexive are identical to each other. Moreover, the reflexive requires the antecedent in a certain syntactic position. This syntactic position can be clarified with Principle A of Binding theory, according to which the reflexive must be bound in a binding category (Chomsky 1981).[1]

[1] The binding category in Principle A refers to following conditions: (i) the antecedent and the reflexive must be co-indexed as well as the antecedent must c-command the reflexive, and (ii) the antecedent must be in a minimal clause which incorporates the reflexive and the accessible SUBJECT (the term referring to the subject in a clause or DP possessor containing anaphora as a potential antecedent).

(2) a. **The new executive** who oversaw the middle managers apparently doubted **himself** on most major decisions.

b. *__The new executive__ who oversaw the middle managers apparently doubted **themselves** on most major decisions.

As in (2) extracted from Dillon et al. (2013), there is the intervening attractor (*the middle managers*) within the dependency relation between the reflexive and its antecedent. While the reader successfully links the reflexive (*himself*) to its antecedent (*the new executive*) as in (2a), the reader does not accept the faulty dependency relation in (2b) due to mismatching features of the reflexive and the antecedent. Meanwhile, despite the matching features of the reflexive (*themselves*) and the attractor (*the middle managers*) in the ungrammatical sentence, it has been found that the reflexive rarely retrieves the attractor in the prior studies (Nicol and Swinney 1989; Sturt 2003; Xiang, Dillon, and Phillips 2009; Dillon 2011; Dillon et al. 2013; Cunnings and Sturt 2014; Parker and Phillips 2017). This result leads to the assumption that the reader retrieves the item based on the syntactic position rather than the feature content in reflexive licensing. In other words, at the retrieval stage in reflexive-antecedent dependencies, syntactic cues, which hold relational information to guide the item to the syntactically licensed position, are privileged rather than morphological cues, which incorporate morphological features to agree between elements. Sentence processing of both subject-verb agreement and reflexive-antecedent dependencies suggests that the reader employs syntactic and morphological cues at the stage of retrieval to directly access a licensor in a content-addressable memory (McElree 2000; McElree, Foraker, and Dyer 2003; Lewis and Vasishth 2005), but the contrasting profiles of attraction in both dependencies imply that there are distinct patterns of cue weightings depending on the types of dependencies.

However, attraction effects in reflexive-antecedent dependencies have often been captured in a few studies (Badecker and Straub 2002; King, Andrews, and Wagers 2012; Cunning and Felser 2013; Patil, Vasishth, and Lewis 2016; Parker and Phillips 2017). One of those studies captured both presence and absence of attraction effects in reflexive processing by manipulating the number of mismatching features between the reflexive and the antecedent (Parker and Phillips 2017). This selective reflexive attraction effects indicate that each retrieval cues are engaged not only in different cue weighting systems but also in different cue combination rules depending on types of dependencies (Parker

2019).

Based on these results in human sentence processing, this study will investigate whether a pre-trained DNN language model, BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018), can show attraction effects of subject-verb agreement and reflexive-antecedent dependencies. This can show us how linguistic cues are combined for dependency formation under the self-attention mechanism of BERT as well as the difference between human and BERT sentence processing. The goal of this research is to enhance explainability and interpretability of DNN language models by understanding the differences in sentence processing mechanisms between human and DNN language models.

2. Background

2.1 Attraction effects in subject-verb agreement

In a content-addressable memory architecture, some pieces of linguistic information, including morphological and syntactic features, are deployed in the memory representation to directly access the target item without searching for the irrelevant representation. The assumptions of a content-addressable memory are provided in several ways. First, when processing sentences with long-distance dependencies, the time course to access the target item is hardly affected by distance between two relational elements for dependency since there is no need to search for every location of irrelevant elements. Instead, it is less likely to retrieve the accurate target item due to memory limitations. That is, the constant retrieval speed and lower comprehension accuracy are construed as evidence of the content-addressable memory (McElree 2000; McElree, Foraker, and Dyer 2003; Lewis and Vasishth 2005). Another assumption is based on an interference effect where a partial feature match gives rise to retrieval of the syntactically irrelevant item rather than the target item (Van Dyke and Lewis 2003; Van Dyke and McElree 2006, 2011; Van Dyke 2007). This effect may induce processing facilitation in ungrammatical sentences during processing a certain type of dependency.

The interference effect has often been observed in comprehension of subject-verb agreement, which captures processing facilitation in ungrammatical sentences with the feature-matching attractors, known as agreement attraction (Pearlmutter, Garnsey, and

Bock 1999; Wagers, Lau, and Phillips 2009; Dillon et al. 2013; Tanner, Nicol, and Brehm 2014; Lago et al. 2015; Tucker, Idrissi, and Almeida 2015; Kim, Brehm, and Yoshida 2019). For example, Wagers et al. (2009) examined the processing of subject-verb agreement with self-paced reading and acceptability judgement. The materials in the experiments included both grammatical and ungrammatical sentences as in (3). In both grammatical and ungrammatical sentences, the number features of attractors were either singular or plural, which led to feature match or mismatch between verb and attractor.

(3) a. **The key** to the cell **was** rusty from many years of disuse.
b. **The key** to the cells **was** rusty from many years of disuse.
c. ***The key** to the cell **were** rusty from many years of disuse.
d. ***The key** to the cells **were** rusty from many years of disuse.
(Wagers et al. 2009: 221)

The experimental results showed that there was no difference found in reading time and acceptability ratings in the grammatical sentences, regardless of feature (mis)match between the verb and the attractor. Instead, in the ungrammatical sentence, the plural attractor caused faster reading time and greater scores on acceptability than the singular one. That is, the syntactically inaccessible but feature-matching attractor led to processing facilitation only in the ungrammatical sentence, showing that agreement attraction interacts with the grammatical asymmetry.

Wagers et al. (2009) claimed that the grammatical-ungrammatical asymmetry in attraction effects during comprehension of subject-verb agreement is due to reanalysis. In this view, if the subject and the verb match on the first-pass, nothing needs to be done. But when the subject and the verb mismatch, the reanalysis process should be engaged to check back to see if there was an error. On the reanalysis stage, the syntactically incorrect but feature-matching attractor might be mistakenly retrieved by adopting morphological features (i.e., number) as retrieval cues, which provides support for a cue-based retrieval mechanism in the content-addressable memory.

2.2 Lack of attraction in reflexive-antecedent dependencies

While comprehension of subject-verb agreement has shown attraction as evidence for the use of morphological features as retrieval cues, comprehension of reflexive-antecedent dependencies have often failed to reflect attraction effects (Nicol and Swinney 1989; Sturt 2003; Xiang, Dillon, and Phillips 2009; Dillon 2011; Dillon et al. 2013; Cunnings and Sturt 2014; Parker and Phillips 2017). These contrasting profiles of attraction between subject-verb agreement and reflexives-antecedent dependencies were found in the study by Dillon et al. (2013). Dillon and his colleagues employed eye-tracking measures to compare comprehension of subject-verb agreement and reflexive-antecedent dependencies as in (4) and (5), in which the verbs (*was/were*) or the reflexives (*himself/themselves*) agree or disagree with the subject (t*he new executive*) in number. Also, there were intervening attractors (*the middle manager(s)*) whose features matched or mismatched the features of the verb or the reflexive.

(4) a. **The new executive** who oversaw the middle manager apparently **was** dishonest about the company's profits.
b. **The new executive** who oversaw the middle managers apparently **was** dishonest about the company's profits.
c. ***The new executive** who oversaw the middle manager apparently **were** dishonest about the company's profits.
d. ***The new executive** who oversaw the middle managers apparently **were** dishonest about the company's profits.

(5) a. **The new executive** who oversaw the middle manager apparently doubted **himself** on most major decisions.
b. **The new executive** who oversaw the middle managers apparently doubted **himself** on most major decisions.
c. ***The new executive** who oversaw the middle manager apparently doubted **themselves** on most major decisions.
d. ***The new executive** who oversaw the middle managers apparently doubted **themselves** on most major decisions.

(Dillon et al. 2013: 89)

In comprehension of both dependencies, the grammatical sentences did not report any difference in reading time measures. Also, during the processing of incorrect subject-verb agreement dependencies like (4c) and (4d), the plural verb (*were*) was susceptible only to the plural attractor (t*he middle managers*), which is consistent with other studies reporting the agreement attraction (earlmutter, Garnsey and Bock 1999; Wagers, Lau, and Phillips 2009; Dillon et al. 2013; Tanner, Nicol, and Brehm 2014; Lago et al. 2015; Tucker, Idrissi, and Almeida 2015; Kim, Brehm, and Yoshida 2019). In contrast, the processing of the ungrammatical reflexive-antecedent dependencies as in (5c) and (5d) did not exhibit attraction effects even when the number feature of the attractor corresponded to that of reflexive, which tells us that the reader is only sensitive to the relation between the reflexive and the syntactically accessible antecedent, not accessing the attractor.

This phenomenon was also true for gender features. In the earlier study, Sturt (2003) revealed, with eye-tracking measure, that reflexive-antecedent dependencies were immune to attraction when the stereotypical gender features were manipulated as in (6). When the gender-biased target noun (*the surgeon*) and the syntactically inaccessible proper noun (*Jonathan/Jennifer*) were potential feature-matching or -mismatching licensors of the reflexive (*himself/herself*), the reader is only responsive to the gender mismatch between the reflexive and the antecedent, independently of the gender features of the attractors.

(6) a. **The surgeon** who treated Jonathan had pricked **himself** with a used syringe needle.
b. **The surgeon** who treated Jennifer had pricked **himself** with a used syringe needle.
c. **The surgeon** who treated Jonathan had pricked **#herself** with a used syringe needle.
d. **The surgeon** who treated Jennifer had pricked **#herself** with a used syringe needle. (Sturt 2003: 555)

The verb and the reflexive in both subject-verb agreement and reflexive-antecedent dependencies superficially seem to target the same structural position (i.e., subject in the sentence). However, the contrasting profiles of attraction show that even though both dependencies use a combination of syntactic and morphological cues to access the target item, the retrieval for reflexive-antecedent dependencies prioritizes syntactic cues over morphological cues.

According to Dillon et al. (2013) and Parker and Phillips (2017), one possibility that can illustrate the different aspects of attraction between subject-verb agreement and reflexive-antecedent dependencies relies on *the predictability of the dependency*. In this view, the subject noun provides predictive information about the number feature of the upcoming verb during the processing of subject-verb agreement. When the feature of the verb disaccords with the prediction, a syntactically incorrect but feature-matching noun is engaged in the cue-based retrieval process at the reanalysis stage. On the other hand, during the processing of reflexive-antecedent dependencies, the subject serving as an antecedent cannot predict the dependency relation until encountering the reflexive. As a result, when encountering the reflexive, the reader preferentially retrieves the item in the syntactically licit position. Therefore, this distinction of the predictability between two dependencies seems to trigger the contrasting profiles of attraction.

2.3 Selective reflexive attraction

While subject-verb agreement processing has consistently reported attraction effects, reflexive-antecedent dependencies processing has not always shown lack of attraction. In fact, a few studies have reported attraction effects in reflexive-antecedent dependencies (Badecker and Straub 2002; King, Andrews, and Wagers 2012; Cunning and Felser 2013; Patil, Vasishth, and Lewis 2016; Parker and Phillips 2017). Among those studies, Parker and Phillips (2017) captured both absence and presence of attraction during reflexive processing with eye-tracking experiments. By slightly changing the degree of feature mismatch (e.g., number, gender and animacy), they robustly elicited processing facilitation from comprehension of reflexive-antecedent dependencies. In the experimental materials as in (7), the degree of feature mismatch between the reflexive and the antecedent was manipulated. In (7a) and (7b), the features of the antecedent (*the studious schoolgirl*) perfectly matched those of the reflexive (*herself*), which meant that the number of feature mismatch was 0 and this sentence was grammatical. Meanwhile, there were ungrammatical sentences with 1-feature mismatch (i.e., gender) between the antecedent (*the studious schoolboy*) and the reflexive (*herself*) as in (7c) and (7d), and 2-feature mismatch (i.e., gender and animacy) between the antecedent (*the brief memo*) and the reflexive (*herself*) as in (7e) and (7f). The 2-feature mismatch designs were replicated in the subsequent experiments in the same study, such as ‘number and gender’ and ‘number and animacy’.

(7) a. The strict librarian said that **the studious schoolgirl** reminded **herself** about the overdue book.
b. The strict father said that **the studious schoolgirl** reminded **herself** about the overdue book.
c. *The strict librarian said that **the studious schoolboy** reminded **herself** about the overdue book.
d. *The strict father said that **the studious schoolboy** reminded **herself** about the overdue book.
e. *The strict librarian said that **the brief memo** reminded **herself** about the overdue book.
f. *The strict father said that **the brief memo** reminded **herself** about the overdue book. (Parker and Phillips 2017: 276)

Along with the previous studies reporting lack of attraction in reflexive processing, attraction was not detected during comprehension of the experimental sentences in the 1-feature mismatch condition. In fact, facilitatory effects were present but too weak to observe and, thus, statistically insignificant. On the other hand, attraction was robustly detected while processing the sentences in the 2-feature mismatch condition. To sum up, during reflexive processing, syntactic cues are weighted more than morphological cues by default (i.e., 1-feature mismatch), but this priority for syntactic cues can be eased, as the degree of feature mismatch between the reflexive and the antecedent gets stronger (i.e., 2-feature mismatch).

To illustrate the cue weighting system, Parker (2019) provided quantitative predictions with computational modeling by suggesting a weighted cue-combinatoric scheme. When the retrieval requires multiple cues to access the memory representation, these cues are combined at the retrieval stage and given strength based on the relationship between retrieval cues and the features in the memory trace (Van Dyke and McElree 2006). In the activation-based model (Lewis and Vasishth 2005), the cue-combinatoric scheme is defined as either a linear cue combination or a non-linear cue combination. The linear cue combination model proposes that each matching cue independently affects the strength of the cue combination, so that the item's activation increases linearly for every matching cue (i.e., additive). On the other hand, the non-linear cue combination model suggests that the strength of the cue-combination is impacted by conjunctions of cues, not by individual cues (i.e., multiplicative).

Parker (2019) applied reading time data to each equation for the linear and non-linear combination rules in order to generate predictions about the likelihood of attraction. As shown in Figure 1, the linear rule expects that the probabilities of retrieving the attractor gradually grow, as the degree of feature mismatch between the reflexive and the antecedent increases. By contrast, the non-linear rule expects that the probabilities of attraction are relatively low at the full match and 1-feature mismatch conditions, but the probability drastically rises at the 2-feature mismatch condition. As the non-linear rule provides a better fit to the previous findings (Parker and Phillips 2017), it seems that selective attraction in reflexive processing occurs in the non-linear fashion. This means that the types of retrieval cues in reflexive-antecedent dependencies are differentially weighted; as expected, syntactic cues are weighted more than morphological cues by default (1-feature mismatch), and the stronger mismatch within dependency (2-feature mismatch) neutralizes the priority for syntactic cues. Meanwhile, in terms of the cue-combinatoric scheme, subject-verb agreement dependencies might be engaged in the linear rule. This assumption is due to the fact that subject-verb agreement can reflect attraction effects even in the 1-feature mismatch condition where the number feature of the subject mismatches that of the verb.

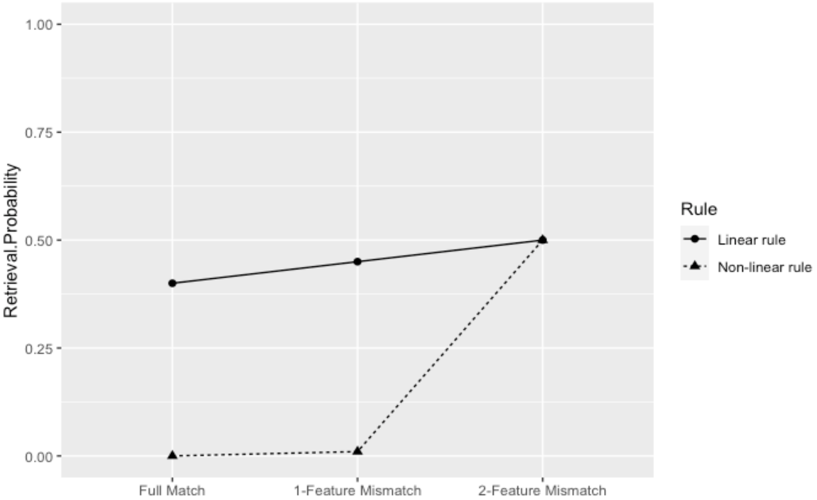

Figure 1. Probability of retrieval for the syntactically incorrect but feature-matching attractor in reflexive-antecedent dependencies based on the degree of the feature mismatch using linear and non-linear cue combination rules (Parker 2019: 20)

2.4 Linguistic representations in BERT

DNN language models are constructed by learning large amounts of corpus data. During the data-training, DNN language models encode numerous words or sentences into vectors (i.e., sequences of real numbers), which are unrelated to classic linguistic information (e.g., morphology, syntax, etc.). These vectors need to go through a series of arithmetic processes in the neural network's internal states to generate human-interpretable output. To understand the DNN language models' internal states, it is common to investigate what kinds of linguistic features are learned and how those features behave (Belinkov and Glass 2019).

BERT is a pre-trained DNN language model based on Transformers (Vaswani et al. 2017), which is a DNN architecture consisting of encoder-decoder layers. In Transformer architecture, when the encoder receives input tokens in a single sentence and vectorizes the tokens, the decoder receives and reconstructs the vectors to generates the output sentence. BERT language model uses only the encoder layers of Transformers and includes a special type of attention mechanism called self-attention, which allows BERT to understand words based on its context. The self-attention relates each token to the other tokens in a single input sequence. Attention heads in each layer denote weighted representations for each input token and the outputs of attention heads are combined and run through the connected layers. As these respective layers in BERT deal with different types of linguistic information, including phrase-level, syntactic and semantic information (Jawahar, Sagot, and Seddah 2019), final outputs are considered as combinations of various pieces of linguistic information.

Among a number of studies proving that the internal representations of BERT language model reflect various types of linguistic information (Goldberg 2019; Hewitt and Manning 2019; Jawahar, Sagot, and Seddah 2019; Lin, Tan, and Frank 2019; Liu et al. 2019; Tenney et al. 2019; Klafka and Ettinger 2020), Lin, Tan, and Frank (2019) showed the evidence that BERT's representations simulate linguistic representations based on hierarchical syntactic structures rather than linear word orders in the long-distance subject-verb agreement and reflexive-antecedent dependencies. In their experiment, the materials were all grammatical sentences in which one or more target nouns and attractors were involved as in (8) and (9). In the materials, the head noun that the verb or the reflexive should be linked to is at the syntactically accessible position based on the hierarchical representation, while one or more attractors are at the linearly closer position to the verb or the reflexive. The result showed that the verb and the reflexive of both dependencies exhibited higher prediction accuracy on the syntactically accessible target noun rather than the linearly closer attractor. This finding suggested that attention weights of BERT reflect the linguistic representations based on syntactic structures, not linear strings.

(8) a. **The cat** near the dog **does** sleep.
b. **The cat** near the dogs **does** sleep.
c. **The cat** that can comfort the dog **does** sleep.
d. **The cat** that can comfort the dogs **does** sleep.
(9) a. **The lord** could comfort the wizard by **himself**.
b. **The lord** could comfort the witch by **himself**.
c. **The lord** that can hurt the prince could comfort **himself**.
d. **The lord** that can hurt the princess could comfort **himself**.
e. **The lord** that can hurt the prince could comfort the wizard by **himself**.
f. **The lord** that can hurt the princess could comfort the wizard by **himself**.
g. **The lord** that can hurt the prince could comfort the witch by **himself**.
h. **The lord** that can hurt the princess could comfort the witch by **himself**.
(Lin et al 2019: 7)

Meanwhile, when they compared baseline sentences involving the feature-mismatching attractor to the sentences involving the feature-matching attractor or two attractors by using confusion scores as a measure, the confusion scores were affected by the feature match and the number of attractors. When there is feature match between the attractor and the verb or the reflexive, or when there is relatively more number of attractors, greater confusion (i.e., processing difficulty) was observed. In correct dependency formation, the greater confusion was attributed to the feature-matching attractor, indicating that attraction effects were present in both subject-verb agreement and reflexive-antecedent dependencies.

Nevertheless, the results of this experiment may not be a reliable indicator of sensitivity to attraction effects, because it only focuses on grammatical sentences. According to observations of human sentence processing, attractors are mostly accessed when processing ungrammatical sentences, rather than grammatical ones. (Nicol and

Swinney 1989; Pearlmutter, Garnsey, and Bock 1999; Sturt 2003; Wagers, Lau, and Phillips 2009; Xiang, Dillon, and Phillips 2009; Dillon 2011; Dillon et al. 2013; Cunnings and Sturt 2014; Tanner, Nicol, and Brehm 2014; Lago et al. 2015; Tucker, Idrissi, and Almeida 2015; Parker and Phillips 2017; Kim, Brehm, and Yoshida 2019). Even though attraction effects have been observed in BERT's language processing, this doesn't necessarily provide evidence that BERT's linguistic representations are based on syntactic structures. In human sentence processing, accessing the feature-matching attractor is viewed as a sign of using morphological information rather than syntactic information. Moreover, Lin et al. (2019) argued that the reflection of reflexive attraction effects in BERT is inconsistent with the observations from human sentence processing (Nicol and Swinney 1989; Sturt 2003; Xiang, Dillon, and Phillips 2009; Dillon 2011; Dillon et al. 2013; Cunnings and Sturt 2014; Parker and Phillips 2017), but a few other studies have observed reflexive attraction effects (Badecker and Straub 2002; King, Andrews, and Wagers 2012; Cunning and Felser 2013; Patil, Vasishth, and Lewis 2016; Parker and Phillips 2017). In particular, Parker and Phillips (2017) showed that the manipulation of cue weightings was the key factor to capture attraction effects in reflexive processing.

Therefore, the present study will investigate whether BERT can accurately capture grammatically asymmetrical attraction effects in subject-verb agreement and reflexive-antecedent dependencies. Additionally, we will examine how linguistic cues are combined during this process. To achieve these goals, we will use a set of experimental materials consisting of both grammatical and ungrammatical sentences, and we will manipulate the degree of feature mismatch in some of these materials. Through this approach, the study seeks to identify how BERT combines linguistic features under the self-attention mechanism, specifically in terms of the cue-combinatoric scheme.

3. Experiment

3.1 Method and procedure

Google AI provides two kinds of pre-trained BERT models; *bert-based-uncased* and *bert-large-uncased. Bert-based-uncased* model has 12 layers, 12 attention heads and 768 hidden units, while *bert-large-uncased* model has 24 layers, 16 attention heads and 1024 hidden units. In this experiment, the pre-trained *bert-base-uncased* model from the HuggingFace's Transformers library (Wolf et al. 2020) is used.[2] Actual codes used in the experiment are provided by Cho et al. (2021).

Pre-training of BERT is implemented in two ways: (i) Masked Language Model (MLM) aims to predict a word for a certain masked area, and (ii) Next Sentence Prediction (NSP) makes decisions on whether the first sentence can be followed by the second sentence, when a pair of sentences is given. By taking advantage of MLM, the current study seeks the probability that a certain word can emerge at a specific position in the sentence. Suppose that the verb position is covered with '[MASK]' as in (10). In this case, we can predict that the verb '*was*', which makes the sentence grammatical, obtains a higher probability at the masked position than the verb '*were*', which leads to the ungrammatical sentence.

(10) The key to the cells [MASK] rusty.

This prediction is brought about by the self-attention mechanism of BERT. *Attention* refers to the extent to which pairs of words are related to each other, and self-attention indicates that the attentions in word embeddings (the sequence of vectors) are applied to themselves (the same strings of the word embeddings). To be specific, when an attention head takes a sequence of word embeddings ($h = [h_1, h_2, \cdots, h_n]$) from the input sentence, each embedding vector consists of queries (Q), keys (K) of dimension d_k, and values (V) of dimension d_v. All pairs of words are computed as softmax-normalized dot products between the queries and the transposed keys, and divided each by $\sqrt{d_k}$. This output is multiplied to a weighted sum of value vectors to obtain attention as in (11).

(11) $$\text{Attention}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Through the softmax layers, each applied attention weight is represented between 0 and 1, and total attention weights in a sequence should be 1. The figure 2 shows the visualizing the self-attention mechanism of BERT with BertViz (Vig 2019). The darkness

[2] The materials, results and codes used in the experiment are provided in: https://github.com/joyennn/Grammatical-illusions-in-BERT

of the line in the figure indicates relatively more weighted attention, which means that two words connected with the darker line are more related with each other than with the other words. For example, in the left panel of Figure 2, the word pair *was* and *key* exhibit the highest attention weight compared to other word pairs with *was*, indicating a strong association between the verb *was* and the subject *key* in the given sentence. Furthermore, based on a comparison of the two panels in Figure 2, it is suggested that the *was*-*key* pair in the left panel is more closely linked than the *were*-*key* pair in the right panel.

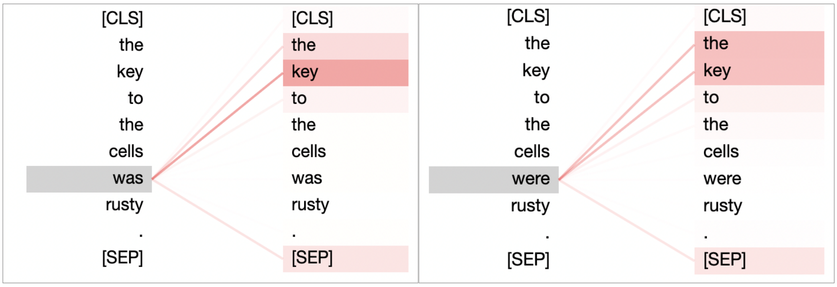

Figure 2. The visualization of the self-attention mechanism with BertViz (Vig 2019)

The probability (P), or the attention weight, for the word emerging in the masked area (w_{mask}) in the given context (*Context*) will be calculated into surprisal (Hale 2001; Levy 2008) as in (12). Surprisal is an effective measure of cognitive effort in language processing in the field of psycholinguistics. A higher surprisal score indicates that the occurrence of the word is less expected, requiring more cognitive effort to process, whereas a lower surprisal score indicates a more natural and expected word occurrence in the sentence. Therefore, surprisal is inversely correlated with the acceptability of a sentence (Hale 2001; Levy 2008).

(12) Surprisal = $-\log_2 P(w_{mask|}Context)$

In this study, surprisal plays a role in representing the level of difficulty in processing sentences, which is similar to a reading time measure used in human sentence processing. Therefore, higher surprisal scores indicate processing inhibition, whereas lower surprisal scores indicate processing facilitation. The surprisal scores obtained from the experiment will be statistically analyzed using linear mixed-effects regression models from the lme4 package in the R statistical software (Bates et al. 2014). Each model will include surprisal scores as fixed effects and items as random intercepts (Baayen et al. 2008).

3.2 Experiment 1A/1B

3.2.1 Materials

Experimental sentences consist of 100 sets of subject-verb agreement dependencies arranged in a 2×2 design, in which 'Grammaticality' (Grammatical / Ungrammatical) and 'Attractor number' (Singular / Plural) are manipulated as shown in Table 1 & 2. Each sentence contains a masked area ('[MASK]') which will be filled with either the verb *was* or *were*. Since the subject in each sentence is singular, the verb *was* makes the sentence grammatical, while the verb *were* makes the sentence ungrammatical. In addition, the number feature of the verb matches or mismatches that of the attractor, which intervenes between the subject and the verb.

In Experiment 1, there are two versions of experimental sentences, as described by Lin et al. (2019). Experiment 1A includes the attractor within a prepositional phrase, while Experiment 1B includes the attractor is within a relative clause. Through these sets of experiments, we will investigate whether agreement attraction effects can be captured.

Table 1. A sample set of experimental items from Experiment 1A

Factor		Item	Verb
Grammaticality	Attractor number		
Grammatical	Singular	The key to the cell [MASK] rusty.	was
Grammatical	Plural	The key to the cells [MASK] rusty.	was
Ungrammatical	Singular	The key to the cell [MASK] rusty.	were
Ungrammatical	Plural	The key to the cells [MASK] rusty.	were

Table 2. A sample set of experimental items from Experiment 1B

Factor		Item	Verb
Grammaticality	Attractor number		
Grammatical	Singular	The editor who rejected the book [MASK] interviewed.	was
Grammatical	Plural	The editor who rejected the books [MASK] interviewed.	was
Ungrammatical	Singular	The editor who rejected the book [MASK] interviewed.	were
Ungrammatical	Plural	The editor who rejected the books [MASK] interviewed.	were

3.2.2 Result

As results of experiment 1A/1B, Figure 3 shows the distributions of the surprisal scores and Table 3&4 present median surprisals obtained from the experimental data. Since there are some outliers that can cause bias in the interpretation of the results, median value is adopted instead of mean. The summaries of applying linear mixed-effects regression models are presented in Table 5&6.

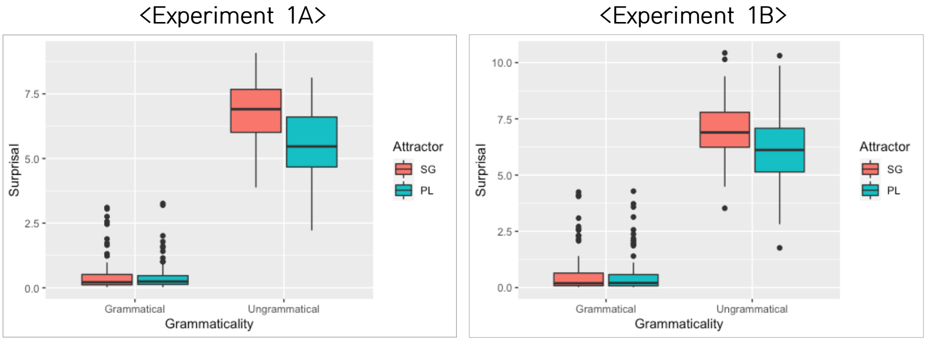
<Experiment 1A>
<Experiment 1B>

Figure 3. The distribution of surprisal scores in Experiment 1A/1B

Table 3. The median surprisals from the result of Experiment 1A

Grammaticality	Grammatical		Ungrammatical	
Median surprisals	0.23		6.35	
Attractor number	Singular	Plural	Singular	Plural
Median surprisals	0.21	0.24	6.9	5.46

Table 4. The median surprisals from the result of Experiment 1B

Grammaticality	Grammatical		Ungrammatical	
Median surprisals	0.2		6.6	
Attractor number	Singular	Plural	Singular	Plural
Median surprisals	0.19	0.2	6.89	6.11

In both experiment 1A and 1B, the surprisal scores in the grammatical sentences (1A: 0.23 / 1B: 0.2) are much lower than those in the ungrammatical sentences (1A: 6.35 / 1B: 6.6). This grammaticality shows statistically significant effects (1A: Estimate = 5.1, SE = 0.14, t = 35.46, p < .001 / 1B: Estimate = 5.54, SE = 0.11, t = 48.34, p < .001). In addition, the median surprisal scores in the grammatical sentences do not show much difference depending on the attractor number (1A: singular attractor = 0.21, plural attractor = 0.24 / 1B: singular attractor = 0.19, plural attractor = 0.2), while the median surprisal scores in the ungrammatical sentences exhibit considerable difference between attractor number features (1A: singular attractor = 6.9, plural attractor = 5.46 / 1B: singular attractor = 6.89, plural attractor = 6.11). Especially, the relatively lower surprisals during the processing of the plural attractor indicate that the feature-matching attractor has an impact on processing facilitation in the ungrammatical sentences. This reflects a significant interaction of grammaticality and attractor number (1A: Estimate = 1.23, SE = 0.2, t = 6.09, *p* < .001 / 1B: Estimate = 0.88, SE = 0.16, t = 5.44, *p* < .001). Nevertheless, no significant effects are observed in attractor number (1A: Estimate = 0.04, SE = 0.14, t = 0.29, *p* = 0.76 / 1B: Estimate = 0.02, SE = 0.16, t = 0.11, *p* = 0.91).

Table 5. Summary of fixed effects from linear mixed-effects models in Experiment 1A

	Estimate	SE	t	*p*-value
(intercept)	0.42	0.10	4.16	**<.001**
Grammaticality	5.1	0.14	35.46	**<.001**
Attractor number	0.04	0.14	0.29	0.76
Grammaticality x Attractor number	1.23	0.20	6.09	**<.001**

Table 6. Summary of fixed effects from linear mixed-effects models in Experiment 1B

	Estimate	SE	t	*p*-value
(intercept)	0.56	0.11	5.01	**<.001**
Grammaticality	5.54	0.11	48.34	**<.001**
Attractor number	0.02	0.16	0.11	0.91
Grammaticality x Attractor number	0.88	0.16	5.44	**<.001**

3.2.3 Discussion

According to the results of Experiment 1A/1B, the ungrammatical sentences show significantly higher surprisal scores than the grammatical sentences. When the attractor numbers differ in the sentences where subject-verb numbers mismatch, the attractors matching the number of the verb show lower surprisal scores than the attractors mismatching the number of the verb, leading to processing facilitation. However, the attractors in the grammatical sentences do not affect either processing inhibition or facilitation. In other words, it is not until the subject mismatches the verb that the feature-matching attractors exhibit processing facilitation. The fact that the attraction occurs only in the ungrammatical sentences tells us that BERT reflects the interaction between grammatical asymmetry and attraction effects during the processing of subject-verb agreement. This result is in accord with the observations from human sentence processing (Pearlmutter, Garnsey, and Bock 1999; Wagers, Lau, and Phillips 2009; Dillon et al. 2013; Tanner, Nicol, and Brehm 2014; Lago et al. 2015; Tucker, Idrissi, and Almeida 2015; Kim, Brehm, and Yoshida 2019).

3.3 Experiment 2A/2B

3.3.1 Materials

Table 7. A sample set of experimental items from Experiment 2A

Factors		Item	Reflexive
Grammaticality	Attractor number		
Grammatical	Singular	The persuasive lawyer that the innocent manager talked about defended [MASK] in the court case.	himself
Grammatical	Plural	The persuasive lawyer that the innocent managers talked about defended [MASK] in the court case.	himself
Ungrammatical	Singular	The persuasive lawyer that the innocent manager talked about defended [MASK] in the court case.	themselves
Ungrammatical	Plural	The persuasive lawyer that the innocent managers talked about defended [MASK] in the court case.	themselves

Table 8. A sample set of experimental items from Experiment 2B

Factors		Item	Reflexive
Grammaticality	Attractor gender		
Grammatical	Masculine	The male lawyer that the male manager talked about defended [MASK] in the court case.	himself
Grammatical	Feminine	The male lawyer that the female manager talked about defended [MASK] in the court case.	himself
Ungrammatical	Masculine	The male lawyer that the male manager talked about defended [MASK] in the court case.	herself
Ungrammatical	Feminine	The male lawyer that the female manager talked about defended [MASK] in the court case.	herself

The second experiment investigates whether attraction effects are captured in reflexive-antecedent dependencies by manipulating number and gender features respectively in two versions of experiments. Each version of the experiments contains 79 sets of experimental materials. In the materials, the antecedents of the reflexives are at the subject position, and the attractors intervene within the relative clauses between the antecedents (i.e., subject) and the reflexives (i.e., object). The reflexive positions are covered with ‘[MASK]’ in the materials.

The materials in the experiment 2A are arranged by a 2×2 design in which ‘Grammaticality’ (Grammatical / Ungrammatical) and ‘Attractor number’ (Singular / Plural) are manipulated as provided in Table 7. The candidates for the masked area are two reflexives with different number features (*himself* and *themselves*). The singular reflexive (*himself*) makes the sentence grammatical while the plural reflexive (*themselves*) does not. The singular or plural number feature of the attractor matches or mismatches the feature of the reflexive.

Likewise, the materials in the experiment 2B are arranged in a 2×2 design with ‘Grammaticality’ (Grammatical / Ungrammatical) and ‘Attractor gender’ (Masculine / Feminine) as suggested in Table 8. Although stereotypical gender features are used in the previous study (Sturt 2003), the current study makes the gender features clearer by adding gendered adjectives (*male* or *female*) to nouns (antecedents and attractors) in case that the gender features of the nouns are perceived as ambiguous. Thus, the reflexive (either *himself* or *herself*) allows for reflexive-antecedent or reflexive-attractor match or mismatch as to gender.

3.3.2 Result

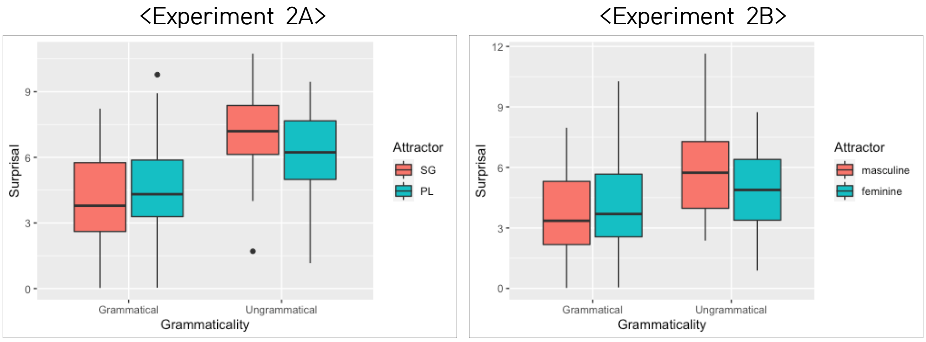

Figure 4. The distribution of surprisal scores in Experiment 2A/2B

Table 9. The median surprisals from the result of Experiment 2A

Grammaticality	Grammatical		Ungrammatical	
Median surprisals	3.98		6.72	
Attractor number	Singular	Plural	Singular	Plural
Median surprisals	3.79	4.32	7.19	6.23

Table 10. The median surprisals from the result of Experiment 2B

Grammaticality	Grammatical		Ungrammatical	
Median surprisals	3.53		5.08	
Attractor gender	Masculine	Feminine	Masculine	Feminine
Median surprisals	3.35	3.69	5.74	4.88

Figure 4 exhibits the distribution of surprisal scores, and Table 9&10 show the median surprisals from the result data of experiment 2A/2B. The summaries of linear mixed-effects regression results are provided in Table 11&12.

In both experiments, the grammatical sentences (2A: 3.98 / 2B: 3.53) have lower surprisals than the ungrammatical sentences (2A: 6.72 / 2B: 5.08). This result reflects significant effects in grammaticality (2A: Estimate = 3.17, SE = 0.16, t = 18.82, *p* < .001 / 2B: Estimate = 2.03, SE = 0.11, t = 17.36, *p* < .001). Besides, surprisals with regard to the attractor feature (number or gender) do not show significant effects (2A: Estimate = 0.42, SE = 0.3, t = 1.37, *p* = 0.17 / 2B: Estimate = 0.42, SE = 0.32, t = 1.31, *p* = 0.19). Also, the difference of surprisals regarding attractor features is significant in the ungrammatical sentences, where the surprisals of the feature-matching attractors (2A: 6.23 / 2B: 4.88) are lower than those of the feature-mismatching attractors (2A: 7.19 / 2B: 5.74). Main effects on the interaction between the grammaticality and the attractor features are observed (2A: Estimate = -1.53, SE = 0.23, t = -6.42, *p* < .001 / 2B: Estimate = -1.23, SE = 0.16, t = -7.47, *p* < .001).

Table 11. Summary of fixed effects from linear mixed-effects models in Experiment 2A

	Estimate	SE	t	*p*-value
(intercept)	3.95	0.21	18.28	**<.001**
Grammaticality	3.17	0.16	18.82	**<.001**
Attractor number	0.42	0.3	1.37	0.17
Grammaticality x Attractor number	-1.53	0.23	-6.42	**<.001**

Table 12. Summary of fixed effects from linear mixed-effects models in Experiment 2B

	Estimate	SE	t	*p*-value
(intercept)	3.64	0.22	15.88	**<.001**
Grammaticality	2.03	0.11	17.36	**<.001**
Attractor gender	0.42	0.32	1.31	0.19
Grammaticality x Attractor gender	-1.23	0.16	-7.47	**<.001**

3.3.3 Discussion

For reflexive-antecedent dependencies, grammaticality and the interaction between grammaticality and attractor feature have significant effects. The ungrammatical sentences show more difficult processing than the grammatical sentences. Moreover, in the ungrammatical sentences, processing facilitation (lower surprisals) is observed for the feature-matching attractors as compared to the feature-mismatching attractors, which indicates that the attraction effects are also captured in reflexive processing. Therefore, both subject-verb agreement and reflexive-antecedent dependencies are processed alike in BERT. This result contrasts with the previous findings that attraction is not observed during reflexive-antecedent dependencies in human sentence processing (Nicol and Swinney 1989; Sturt 2003; Xiang, Dillon, and Phillips 2009; Dillon 2011; Dillon et al. 2013; Cunnings and Sturt 2014; Parker and Phillips 2017).

3.4 Experiment 3

3.4.1 Materials

Table 13. A sample set of experimental items for Experiment 3

Factors		Item	Reflexive
The number of feature mismatch (Grammaticality)	Attractor match		
0 (Grammatical)	MM (Feminine)	The male lawyer that the female manager mentioned defended [MASK] in the court case.	himself
0 (Grammatical)	M (Masculine)	The male lawyer that the male manager mentioned defended [MASK] in the court case.	himself
1 (Ungrammatical)	MM (Masculine)	The male lawyer that the male manager mentioned defended [MASK] in the court case.	herself
1	M	The male lawyer that the female manager mentioned	herself

(Ungrammatical)	(Feminine)	defended [MASK] in the court case.	
2 (Ungrammatical)	MM (Masculine)	The male lawyers that the male manager mentioned defended [MASK] in the court case.	herself
2 (Ungrammatical)	M (Feminine)	The male lawyers that the female manager mentioned defended [MASK] in the court case.	herself

Experiment 3 is designed for the question of whether the degree of cue combination in reflexive-antecedent dependencies has an impact on attraction effects. The materials are composed of 79 sentence sets with a 3×2 design containing ‘The number of feature mismatch’ between the antecedent and the reflexive (0, 1, and 2) and ‘Attractor match’ to the reflexive (mismatch - MM, match – M). When the number of feature mismatch is 0, the feature of the antecedent (*the male photographer*) is perfectly identical to the feature of the reflexive (*himself*), such as number and gender. When the number of feature mismatch is 1, the gender features of the antecedent and the reflexive are not identical to each other, while the number feature of the antecedent is still same as that of the reflexive. When the number of feature mismatch is 2, neither gender nor number features is identical between the antecedent and the reflexive. That is, the number of feature mismatch indicates the degree of grammaticality. Also, the gender feature of the attractor is manipulated for match or mismatch to the feature of the reflexive.

3.4.2 Result

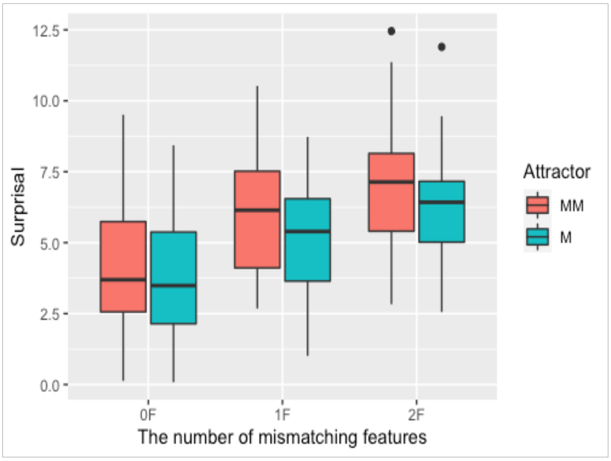

Figure 5. The distribution of surprisal scores in Experiment 3

Table 14. The median surprisals from the result of Experiment 3

The number of feature mismatch	0		1		2	
Median surprisals	3.63		5.81		6.71	
Attractor match	MM	M	MM	M	MM	M
Median surprisals	3.69	3.49	6.14	5.39	7.14	6.42

The figure 5 provides the boxplot graph regarding the distribution of surprisals. Table 14 provides median surprisal scores. Table 15 provides a summary of the linear mixed-effects analysis.

Main effects are observed in the number of mismatching features (i.e., the degree of grammaticality) for both 1-feature (Estimate = 2.00, SE = 0.11, t = 17.47, p < .001) and 2-feature mismatch conditions (Estimate = 3.33, SE = 0.28, t = 11.56, p < .001). This is driven by higher surprisals in the ungrammatical sentences (1-feature = 5.81 / 2-feature = 6.71) relative to the grammatical sentences (0-feature = 3.63). Moreover, there are significant interactions of grammaticality with attractor match in 1-feature (Estimate = -1.00, SE = 0.16, t = -6.18, p < .001) and 2-feature mismatch conditions (Estimate = -1.13, SE = 0.40, t = -2.80, p < .001). 1. Ungrammatical sentences with feature-matching attractors show a more significant effect, resulting in relatively lower surprisal values (1-feature = 5.39 / 2-feature = 6.42), than those with feature-mismatching attractors (1-feature = 6.14 / 2-feature = 7.14). However, no significant effects are observed in attractor match (Estimate = 0.35, SE = 0.27, t = -1.29, p = 0.19).

Table 15. Summary of fixed effects from linear mixed-effects models in Experiment 3

	Estimate	SE	t	*p*-value
(intercept)	3.71	0.19	18.92	**<.001**
1-Feature	2.00	0.11	17.47	**<.001**
2-Feature	3.33	0.28	11.56	**<.001**
Attractor match	0.35	0.27	1.29	0.19
1-Feature x Attractor match	-1.00	0.16	-6.18	**<.001**
2-Feature x Attractor match	-1.13	0.40	-2.80	**<.001**

3.4.3 Discussion

The experiment 3 examines the relation between the degree of reflexive-antecedent feature mismatch and attraction effects. Overall, whether the reflexive mismatches the antecedent in gender (1-feature mismatch) or gender and number (2-feature mismatch), attraction effects are observed. These results demonstrate that the reflexive is susceptible to attraction regardless of the degree of feature mismatch between the antecedent and the reflexive.

However, despite the results in experiment 3, it is still unknown whether the stronger feature mismatch increase probabilities of attraction effects as in human sentence processing. Therefore, the results are applied to the cue-combinatoric scheme (Parker 2019), which represents the relation of cue combination and attraction effects in reflexive-antecedent dependencies. As shown in Figure 1, the probabilities of retrieving the feature-matching attractors during the reflexive processing can increase, as the degree of feature mismatch grows. This additive growth in the probabilities of attraction is regarded as the linear cue-combinatoric rule. On the other hand, the probabilities of retrieving the feature-matching attractors show little increase when the degree of feature mismatch is less strong (i.e., 0 and 1-feature mismatch conditions) but the probability robustly grows when the degree of feature mismatch get stronger (i.e., 2-feature mismatch condition). In human sentence processing, the probabilities of attraction in reflexive-antecedent dependencies follow this multiplicative growth, that is, the non-linear cue-combinatoric rule. On the contrary, Figure 6 illustrates that, in BERT-based sentence processing, the probabilities of retrieving feature-matching attractors increases with the number of feature mismatch between the antecedent and the reflexive. This increasing pattern of probabilities of attraction means that each matching cue additively affects the retrieval of the attractor, following the linear cue-combinatoric rule.

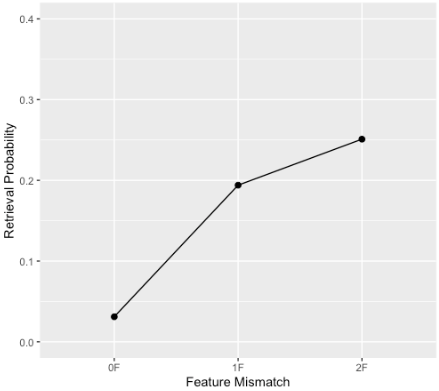

Figure 6. Probability of retrieval for the syntactically incorrect but feature-matching attractor in reflexive-antecedent dependencies from the result in Experiment 3

4. General discussion

This study aims to investigate how BERT, a DNN language model, accesses and combines different types of linguistic information as retrieval cues in processing subject-verb agreement and reflexive-antecedent dependencies. These research questions are motivated by observations of how humans process and comprehend sentences that involve subject-verb agreement and reflexive-antecedent dependencies (Sturt 2003; Wagers, Lau, and Phillips 2009; Dillon et al. 2013; Parker and Phillips 2017). This provides us with insight into how DNN language models process language internally. Through three sets of experiments using MLM in BERT, the study discovered that the mismatch of the subject and the verb as well as the antecedent and the reflexive enables access to the grammatically irrelevant but feature-matching attractor. In addition, more mismatching features in reflexive processing result in stronger attraction effects. In terms of the cue-combinatoric scheme, the combination of different types of retrieval cues follows the linear cue-combinatoric scheme in both dependencies.

In human sentence processing, attraction effects have been consistently observed in subject-verb agreement dependencies. Reflexive-antecedent dependencies, on the other hand, have only selectively shown such effects. However, this study found that both types of dependencies are susceptible to attraction effects in BERT sentence processing. Specifically, experiment 1A/1B aligns with previous research by demonstrating that attraction occurs only in ungrammatical subject-verb agreement conditions containing feature-matching attractors (Wagers, Lau, and Phillips 2009). In contrast, the findings from experiment 2A/2B diverge from previous studies that reported a lack of attraction in reflexive processing (Sturt 2003; Dillon et al. 2013; Parker and Phillips 2017). Therefore, the findings of this study suggest that there are differences between BERT sentence processing and human sentence processing of subject-verb agreement and reflexive-antecedent dependencies. Hence, the result from experiment 3 also cannot follow the observation from the prior study (Parker and Phillips 2017) that exhibited the lack of attraction in the 1-feature mismatch condition while revealing the presence of attraction in the 2-feature mismatch condition. In sentence processing of BERT, attraction effects are observed even in the 1-feature mismatch condition before manipulating the degree of feature mismatch as well as in the 2-feature mismatch condition. Furthermore, these results show that attraction effects in reflexive processing strengthen, as the number of mismatching features increases, following the linear cue-combinatoric scheme. These results also differ from previous findings that reflexive-antecedent dependencies follows the non-linear cue-combinatoric scheme, where the 1-feature mismatch condition rarely captured attraction effects while the 2-feature mismatch condition robustly elicited attraction effects (Parker and Phillips 2017; Parker 2019). In terms of the cue-based retrieval mechanism, the findings from this study suggests that subject-verb agreement and reflexive-antecedent dependencies both use a combination of morphological and syntactic cues without preferring one to the other.

Then, why do human sentence processing and BERT show different patterns regarding attraction effects? Based on the analysis, it is hypothesized that the self-attention mechanism of BERT may not effectively capture the predictability of the dependency. In human sentence processing, subject-verb agreement dependencies are commonly considered to be predictable. Specifically, the subject enables the reader to predict information about the verb before encountering the verb. When the feature of the verb unexpectedly mismatches that of subject, the reader needs to reanalyze the sentence in order to resolve the mismatch. In the reanalysis process, the reader might directly retrieve the grammatically illicit but feature-matching noun, reflecting attraction effects. In contrast, reflexive-antecedent dependencies are typically considered less predictable, as the antecedent does not make a prediction about the upcoming reflexive. Thus, only when encountering the reflexive, can the reader access the antecedent that is syntactically suitable for the reflexive, reflecting the lack of attraction effects.

On the other hand, the self-attention mechanism of BERT integrates the reanalysis process of human sentence processing into a single step. During sentence processing, attention heads in BERT consider every possible relation with the input sequence, linking elements in the dependency to not only the relevant elements but also the irrelevant ones, including the attractor. This allows even the reflexive in BERT to access the attractor as if it had participated in the reanalysis process. Additionally, BERT has shown that the strength of attraction effects is related to the degree of feature mismatch between the antecedent and reflexive, with stronger effects occurring in cases of 2-feature mismatch. Since the sum of all weights in the sequence should equal 1, the decreased attention weight for one word in the 2-feature mismatch condition may lead to an increased attention weight for the attractor, resulting in stronger attraction effects. In summary, BERT shows sensitivity to attraction effects in reflexive processing, and a poorer match to the retrieval cues results in stronger attraction.

While this study found evidence for attraction effects in both subject-verb agreement and reflexive-antecedent dependencies, it is unclear whether the self-attention mechanism in BERT can fully explain these findings within the context of cue-based retrieval. In human sentence processing, cue-based retrieval involves directly accessing the best-matching cues for an item, without considering irrelevant elements. In contrast, the self-attention mechanism in BERT considers all elements, including irrelevant ones, when forming dependencies. In this regard, the retrieval mechanism in BERT might be more closely related to a search mechanism (McElree and Dosher 1993; Gronlund, Edwards, and Ohrt 1997; McElree 2001, 2006), since both the search mechanism and the self-attention mechanism systematically evaluate intervening items within a dependency relation. However, this approach also poses problems. While the search mechanism in human sentence processing can prevent interference effects, the self-attention mechanism in BERT is susceptible to such effects. Therefore, neither the cue-based retrieval mechanism nor the search mechanism seem to fully account for the self-attention mechanism in BERT.

While the self-attention mechanism used in BERT cannot be fully explained by any memory retrieval mechanisms, it is worth considering whether modified self-attention mechanisms can be applied in the context of human sentence processing. Many attempts have been made to modify the self-attention mechanism to attend to only the relevant parts of the sentence since this mechanism requires time and memory that increase quadratically with the sequence length. To mitigate these limitations of the self-attention mechanism, researchers have investigated efficient forms of attention that selectively focus on the related words to improve performance and interpretability of Transformers (Child et al. 2019; Correia et al. 2019; Dai et al. 2019; Roy et al. 2021). For instance, Child et al. (2019) introduced the Sparse Transformer, which learns a fixed pattern of attention connections between input sequence tokens during training to reduce computational complexity. Dai et al. (2019) proposed the Adaptive Attention Span in Transformers, which dynamically adjusts the attention span of the self-attention mechanism based on the importance of each input token. Correia et al. (2019) presented a hierarchical co-attention mechanism that computes attention scores between image and question features at multiple levels of abstraction to compute weighted feature representations. Roy et al. (2021) integrated a gated recurrent unit (GRU) to attend to different parts of the input sequence, enabling the model to attend to specific positions without attending to all positions as in the standard Transformer.

However, while these selective attention mechanisms have been shown to improve the computation time and accuracy of long-distance dependencies in many modified Transformer models, they may not be sufficient for capturing the different patterns of attraction effects in these contexts. This is because selective attention mechanisms focus on attending to specific parts of the input, rather than taking into account the predictability of the dependency. These significant differences between how BERT processes language and how humans process language raise questions about the aspects of human sentence processing that are still not fully understood by BERT.

The different approaches of BERT and human sentence processing to subject-verb agreement and reflexive-antecedent dependencies offer valuable insights for comparing DNN language models and human language processing. Specifically, humans flexibly adapt retrieval mechanisms and prioritize cues depending on the types of dependencies, whereas BERT conforms to persistent ways of retrieval and cue weightings regardless of the types of dependencies. Therefore, a potentially interesting future research could explore whether encoding information about the predictability of the dependency could improve performance on human-like language processing tasks. Additionally, a single theory on memory retrieval mechanism used by human sentence processing might not fully account for the self-attention mechanism used by BERT. Although this study focuses on a particular language model, there are other DNN language models that also use the self-attention mechanism found in Transformers. Therefore, it could be useful to explore if other models that use similar mechanisms to BERT might be more similar to the way humans retrieve memories. Addressing these differences could lead to the development of more human-like language models with broader applicability in natural language processing research.

Although this study has provided valuable insights, there are some limitations that must be addressed in future research. While a larger number of experimental materials are used in this experiment than in typical human sentence processing experiments, the quantity is still not adequate to produce satisfactory results in natural language processing (NLP) experiment. Future research should aim to increase the quantity of materials used in experiments using DNN language models to improve the reliability and validity of the results. Furthermore, despite the contribution of this study to our understanding of BERT and human language processing in the specific tasks and contexts, these findings may generalize to other types of language processing tasks or contexts with caution. Future research should evaluate the generalizability of these findings in broader settings.

Despite some limitations, the present study offers valuable insights into how DNN language models process language compared to human sentence processing. By highlighting these differences, this study contributes to our understanding of language processing and could inspire further research in this area.

5. Conclusion

The current study examines how different types of linguistic information are accessed and combined in the linguistic representations of BERT in comparison with observations from human sentence processing. Across three sets of experiments, our results demonstrate that both subject-verb agreement and reflexive-antecedent dependencies are susceptible to attraction effects. Furthermore, the reflexive attraction reflects the linear cue-combinatorics rule that additively uses a combination of morphological and syntactic cues. Finally, this study proposes that the differences between attraction effects in human sentence processing and BERT may be due to the predictability of the dependency. Overall, these findings can contribute to a deeper understanding of the underlying mechanisms of language processing in humans and language models like BERT. The implications of these findings are significant for the development of xAI systems that aim to better capture the complexities of human language processing and provide more understandable and interpretable explanations for the outputs. However, further research is needed to explore the generalizability of our findings to a broader range of language processing tasks and contexts, and to investigate the implications of our results for the development of future language models.

References

Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4): 390-412.

Bacon, Geoff and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv:1908.09892.*

Badecker, William and Kathleen Straub. 2002. The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(4): 748-769.

Bates, Douglas, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1-48

Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics* 7: 49-72.

Bock, Kathryn and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology* 23(1): 45-93.

Child, Rewon, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv*:1904.10509.

Cho, Won Ik, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the influence of verb aspect on the activation of typical event locations with BERT. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 2922-2929. Association for Computational Linguistics.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Correia, Gonçalo M., Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 2174-2184.

Cunnings, Ian and Claudia Felser. 2013. The role of working memory in the processing of reflexives. *Language and Cognitive Processes* 28(1-2): 188-219.

Cunnings, Ian and Patrick Sturt. 2014. Coargumenthood and the processing of reflexives. *Journal of Memory and Language* 75: 117-139.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Rusland Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 2978-2988.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies,* 4171-4186. Association for Computational Linguistics.

Dillon, Brian William. 2011. *Structured access in sentence comprehension*. PhD Dissertation. University of Maryland.

Dillon, Brian, Alan Mishler, Shayne Sloggett, and Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language* 69(2): 85-103.

Goldberg, Yoav. 2019. Assessing BERT's syntactic abilities. *arXiv:1901.05287.*

Gronlund, Scott D., Mark B. Edwards, and Daryl D. Ohrt. 1997. Comparison of the retrieval of item versus spatial position information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(5): 1261-1274.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies,* 1195-1205. Association for Computational Linguistics.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics,* 159-166. Association for Computational Linguistics.

Hewitt, John and Christopher D. Manning. 2019. 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),* 4129-4138. Association for Computational Linguistics.

Holzinger, Andreas, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9(4): e1312.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language?. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651-3657. Association for Computational Linguistics.

Kim, Nayoun, Laurel Brehm, and Masaya Yoshida. 2019. The online processing of noun phrase ellipsis and mechanisms of antecedent retrieval. *Language, Cognition and Neuroscience* 34(2): 190-213.

King, Joseph, Caroline Andrews, and Matthew Wagers. 2012. Do reflexives always find a grammatical antecedent for themselves. Poster presented at *the 25th Annual Meeting of the CUNY Conference on Human Sentence Processing.* New York, NY: The CUNY Graduate Center. March 14-16.

Klafka, Josef and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4801-4811.

Kush, Dave W. 2013. *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing*. PhD Dissertation. University of Maryland.

Lago, Sol, Diego E. Shalom, Mariano Sigman, Ellen F. Lau, and Colin Phillips. 2015. Agreement attraction in Spanish comprehension. *Journal of Memory and Language* 82: 133-149.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3): 1126-1177.

Lewis, Richard L. and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29(3): 375-419.

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: getting inside BERT's linguistic knowledge. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* 241-253.

Linzen, Tal, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4: 521-535.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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), 1073-1094. Association for Computational Linguistics.

Marvin, Rebecca and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192-1202.

McElree, Brian. 2000. Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research* 29(2): 111-123.

McElree, Brian. 2001. Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(3): 817.

McElree, Brian. 2006. Accessing recent events. *Psychology of Learning and Motivation* 46: 155-200.

McElree, Brian and Barbara A. Dosher. 1993. Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General* 122(3): 291-315.

McElree, Brian, Stephani Foraker, and Lisbeth Dyer. 2003. Memory structures that subserve sentence comprehension. *Journal of Memory and Language* 48(1): 67-91.

Nicol, Janet and David Swinney. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research* 18(1): 5-19.

Parker, Dan and Colin Phillips. 2017. Reflexive attraction in comprehension is selective. *Journal of Memory and Language* 94: 272-290.

Parker, Dan. 2019. Cue combinatorics in memory retrieval for anaphora. *Cognitive Science* 43(3), e12715.

Patil, Umesh, Shravan Vasishth, and Richard L. Lewis. 2016. Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology* 7: 329.

Pearl, Judea. 2018. *The book of why: The new science of cause and effect.* New York City: Basic Books.

Pearlmutter, Neal J., Susan M. Garnsey, and Kathryn Bock. 1999. Agreement processes in sentence comprehension. *Journal of Memory and Language* 41(3): 427-456.

Perconti, Pietro and Alessio Plebe. 2020. Deep learning and cognitive science. *Cognition* 203: 104365.

Phillips, Colin, Matthew W. Wagers, and Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (ed.), *Experiments at the interfaces*, 147-180. Bingley: Emerald Group Publishing Limited.

Roy, Aurko, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics* 9: 53-68.

Sturt, Patrick. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language* 48(3): 542-562.

Tanner, Darren, Janet Nicol, and Laurel Brehm. 2014. The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language* 76: 195-215.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. Presented at *2019 International Conference on Learning Representations (ICLR 2019) Debugging Machine Learning Models Workshop*. New Orleans, LA. May 6-9.

Tucker, Matthew A., Ali Idrissi, and Diogo Almeida. 2015. Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology* 6: 347.

Van Dyke, Julie A. 2007. Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(2): 407-430.

Van Dyke, Julie A. and Richard L. Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language* 49(3): 285-316.

Van Dyke, Julie A. and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language* 55(2): 157-166.

Van Dyke, Julie A. and Brian McElree. 2011. Cue-dependent interference in comprehension. *Journal of Memory and Language* 65(3): 247-263.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Proceedings of the Neural Information Processing Systems(NIPS)*, 5999-6009.

Vig, Jesse. 2019. BertViz: A tool for visualizing multihead self-attention in the BERT model. Presented at *2019 International Conference on Learning Representations (ICLR 2019) Debugging Machine Learning Models Workshop*. New Orleans, LA. May 6-9.

Wagers, Matthew W., Ellen F. Lau, and Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61(2): 206-237.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 38-45.

Xiang, Ming, Brian Dillon, and Colin Phillips. 2009. Illusory licensing effects across dependency types: ERP evidence. *Brain and Language* 108(1): 40-55.

Ye-eun Cho
Graduate student
Department of English Language and Literature

Sungkyunkwan University
25-2, Sungkyunkwan-ro, Jongno-gu
Seoul, 03063 Korea
E-mail: joyenn@skku.edu

Received: 2023. 02. 23.
Revised: 2023. 05. 16.
Accepted: 2023. 05. 23.