https://doi.org/10.24303/lakdoi.2021.29.1.183

Is c-command Machine-learnable?[*]

Unsub Shin, Myung-Kwan Park & Sanghoun Song[**]
(Korea University & Dongguk University)

Shin, Unsub; Park, Myung-Kwan & Song, Sanghoun. (2021). Is c-command machine-learnable? *The Linguistic Association of Korea Journal*, *29*(1), 183-204. Many psycholinguistic studies have tested whether pronouns and polarity items elicit additional processing cost when they are not c-commanded. The previous studies claim that the c-command constraint regulates the distribution of relevant syntactic objects. As such, the syntactic effects of the c-command relation are greatly affected by the types of licensing (e.g. quantificational binding) and reading comprehension patterns of subjects (e.g. linguistic illusion). The present study investigates the reading behavior of the language model BERT when the syntactic processing of relational information (i.e. X c-commands Y) is required. Specifically, our two experiments contrasted the BERT comprehension of a c-commanding licensor versus a non-c-commanding licensor with reflexive anaphora and negative polarity items. The analysis based on the information-theoretic measure of surprisal suggests that violations of the c-command constraint are unexpected for BERT representations. We conclude that deep learning models like BERT can learn the syntactic c-command restriction at least with respect to reflexive anaphors and negative polarity items. At the same time, BERT appeared to have some limitations in its flexibility to apply compensatory pragmatic reasoning when a non-c-commanding licensor intruded in the dependency structure.

Key Words: c-command, deep learning, BERT, surprisal, NPI, reflexive anaphor

[*] This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03031616).

[**] The first author is Unsub Shin. Myung-Kwan Park and Sanghoun Song are the co-corresponding authors.

1. Introduction

Generative grammar has refuted the claim that usage-based stochastic algorithms can learn the hierarchically structured linguistic representations (Everaert et al., 2015). However, the recent empirical studies have accumulated evidence that neural probabilistic language models (also known as distributional word embedding models) are able to capture the implicit structure of natural language by extracting distributional similarities of linguistic expressions (Clark et al., 2019; Hewitt & Manning, 2019; Lin et al., 2019). In addition, analytic evaluations have shown that neural probabilistic language models are often accurate in representing the linguistic property such as number, gender and negation features (Marvin & Linzen, 2018; Futrell et al., 2019; Warstadt et al., 2019).

In keeping with the recent works, we further examine whether the neural language model can learn the implicit hierarchical relation (i.e. c-command). In retrieving the syntactic representation from the memory of deep learning (i.e. feature vectors), we leverage long-distance dependency formations of reflexive anaphors and negative polarity items (NPIs) spanning over intervening materials (i.e. attractors). Previous psycholinguistic studies carrying out the analogous experiment additionally embedded a non-c-commanding attractor that only partially matches some semantic features. Consider the italicised attractors in sentences like (1), which are adapted from Xiang et al. (2009).

(1) a. *The restaurants that *no* local newspapers have recommended in their dining reviews have <u>ever</u> gone out of busine

b. *The tough soldier that *Katie*$_i$ treated in the military hospital introduced <u>herself</u>$_i$ to all the nurses

As such, embedded attractors cannot fully license the use of relevant lexical items in principle. This interference paradigm has been employed as a diagnostic method in order to analyze how language processors access linguistic representations in memory (Xiang et al., 2009; Dillon et al., 2013; Kush et al., 2015; Parker & Phillips, 2016).

Inspired by the presented psycholinguistic approach to encoded linguistic representations, we conducted two language experiments in order to probe the language processing pattern of pre-trained BERT (Devlin et al., 2019). BERT is simply trained to estimate the most likely word in a given context when some word tokens are randomly masked (i.e. self-supervision). The first experiment focused on the BERT comprehension of a c-commanding antecedent vs. a non-c-commanding antecedent when reflexive anaphors are presented under the 2 × 2 factorial design. The second experiment observed the BERT understanding of a c-commanding NPI licensor vs. a non-c-commanding NPI licensor when polarity items must be licensed in a negative context. For the analysis of the language model response, we employed logarithmic surprisal (Levy, 2008) to compute the degree of ill-formedness when the structural condition is not met.

This article is structured as follows. Section 2 illustrates background information of BERT and language experiments relevant to the present study. Section 3 concerns the experimental methods and materials for our empirical investigation of BERT's syntactic processing of the c-command condition. Section 4 provides a statistical analysis and error analysis of BERT acceptability judgments on reflexive anaphors and NPIs. Section 5 suggests fundamentals of how BERT represents the underlying syntactic operations and how BERT diverges from human processors. Section 6 concludes the current data-oriented study with some thoughts of future work.

2. Background

2.1. C-command Condition

Structural c(onstituent)-command is a formal requirement in licensing negative polarity items (NPIs) and in binding reflexive anaphors (Klima, 1964; Ladusaw, 1979; Chomsky, 1981; Kim, 2010). Typically, c-command relation indicates a certain hierarchical constraint which can be defined as follows:

(i) A branching node X that dominates a node A also dominates a node B.
(ii) The nodes A and B do not dominate each other.

Deduced from the theorem, the number feature of the c-commanded pronoun *herself* in (2a) must be dependent on that of the c-commanding binder *the girl* rather than that of the sequentially adjacent plural DP *the servants*. If the reflexive pronoun agrees in number (and in referential index) with the structurally irrelevant antecedent *servants*, then the given sentence is ill-formed like (2b).

(2) a. The girl that hates the servants$_i$ speaks with *herself*$_i$
b. *The girl that hates the servants$_i$ speaks with *themselves*$_i$

To clarify the bound variable reading of the c-commanded reflexive anaphor, consider the example from Reinhart & Reuland (1993). In the example (3), the reflexive pronoun *herself* is bound to the referentially covarying (i.e. coindexed) *Lucie*. This is supported by the logical analysis that (3a) is equivalent to (3b). The constituency test in (3c) likewise proves that *herself* is a bound-variable pronoun as the coordinated clause *Lili (did) too* can be unambiguously replaced by *Lili praised Lili*. This binding relation is met via the agreement condition between the binding antecedent DP and the bound-variable pronoun.

(3) a. Lucie$_i$ praised herself$_i$
b. Lucie (λx (x praised x))
c. Lucie$_i$ praised herself$_i$, and Lili (did) too

The same structural dependency can be found in the NPI licensing condition. Under the analysis of Klima (1964) and Ladusaw (1979), NPIs such as *ever* and *any* are semantically (via downward entailment) licensed in the c-commanded domain of the licensor that constitutes a negative context. Consider the example (4). An NPI *ever* in the example (4b) is not licensed due to the absence of a negative context. In addition, the sentence (4c) shows that the linearly preceding negative quantifier *no* cannot license an NPI *ever* when the licensor is deeply embedded in a subject relative clause.

(4) a. *No*$_i$ boy who had a toy was *ever*$_i$ anxious
b. *The boy who had a toy was *ever*$_i$ anxious
c. *The boy who had *no* toy was *ever*$_i$ anxious

These linguistic data shows that an NPI requires negative context-forming semantic operators to occur in the structurally c-commanding position. But the syntactic requirement for NPI licensing might be less straightforward than reflexive binding, as NPIs have been alternatively accounted for by resorting to other notions such as semantic entailment and pragmatic scalar implicature (Chierchia, 2013).

2.2. Neural Language Models

Deep neural language models like BERT (Devlin et al., 2019) are trained to predict a masked word in a given context using a specific type of deep learning architectures and representational strategies. The BERT language model used the encoder of Transformer (Vaswani et al., 2017) architecture and transfer learning methodology from unlabeled texts. Importantly, BERT integrates the adjoining left and right context of a masked word by simulating the Cloze completion task (Taylor, 1953). This linguistically motivated model training had achieved the state-of-the-art performance in many downstream tasks.

The original Transformer architecture followed the encoder-decoder arrangement of deep neural networks. The encoder first reads an input sequence of words and compresses it into a lower dimensional vector (i.e. distributed representation) through linear transformations. The decoder then represents an output sequence of words by reconstructing the vectorized text information (i.e. *context vector*). But BERT uses only the encoder part of Transformer that functions as a mapping of linguistic feature vectors and their weighted average (i.e. *attention mechanism*) over some sequences of words fed to the model.

In addition, BERT employs the transfer learning methodology which generally combines multiple sources of training corpora. Although these language resources do not share the same probabilistic distribution of data, empirical studies under transfer learning paradigms proved that such corpora as books and news articles share some common semantic and syntactic structures. Under the empirical assumption, BERT is initially pre-trained over multiple text corpora using masked language modelling and then fined-tuned to the downstream tasks. The masked language modelling imitates “Cloze procedure” that measures potential readability of a targeted word/phrase in both top-to-down and bottom-to-up fashions.

Following the Cloze method, BERT embeds the particular [MASK] token at random and predicts the most probable answer that can convey the meaning of the original sentence. Consider the schematic example (5). The denotation [CLS] is a sign of the special token that occupies the first position of every sequence. The [CLS] representation stores aggregate information over the whole sentence and then funnels it into the output layer for the downstream tasks such as a document classification.

(5) a. [CLS] He left the *house* without money. (unchanged token)
b. [CLS] He left the [MASK] without money. ([Mask] token)
c. [CLS] He left the *barn* without money. (random token)

As the example (5) illustrates, BERT first automatically generates the ‘corrupted’ sentence (5b) from the original input sentence (5a). BERT then calculates the probability of the occurrence of *house* in the given context, incorporating the left context *He left the* … and the right context … *without money*. Note that BERT swaps only one token [MASK] in the position of the input word *house*.

For the schematic illustration of an internal operation of the masked language modelling using attention mechanism, suppose the model returned a wrong answer like *barn* as in the sentence (6c) other than *house*. Then the machine updates some parameters relevant to the failed prediction (i.e. linguistic features) and their weights (i.e. the relative importance of intervening linguistic features). It follows that BERT approximates the most likely answer by learning how a masked word is dependent on the other adjacent words from automatically generated probabilistic expectation. But the availability of encoded linguistic information is highly undermined since the black-box system of deep learning only returns the final output of the model training while the internal operation of deep learning remains unidentified.

Accordingly, several probing methods have been proposed to assess the internal “black-box” representation of BERT embeddings and its language modelling method. Hewitt & Manning (2019) used transformation matrices (i.e. *structural probe*) to visualize the encoded non-linear syntactic distance between words in tree structures. In addition, probing models of Liu et al. (2019) found empirical evidence that the bidirectional BERT language model had robustly encoded transferable syntactic information such as part-of-speech tags and boundaries of named entities.

However, recent empirical studies have found that BERT is still a sub-optimal language model that demands more training data and longer training time for achieving high accuracy. For instance, RoBERTa (Liu et al., 2019) have showed that the exclusion of the unsupervised next sentence prediction task does not harm the accuracy of BERT while doubling pre-training data enhances the performance of the model. Moreover, POWER-BERT (Goyal et al., 2020) have showed that the inference time of BERT can be significantly reduced by eliminating redundant linguistic representations without damaging the accuracy. Despite the fact that BERT is not robustly enough optimized, we adopt the standard BERT model since it shares the general architecture and training methodology of BERT with the state-of-the-art models.

2.3. Language Experiments

Although theoretical linguistic data validate some hierarchical constraints, many psycholinguistic studies have further assessed the relative strength of the c-command constraint using the human reading comprehension tasks. Cunnings et al. (2015) and Kush et al. (2015) probed whether c-commanding QP (e.g. no girl scout) is more easily accessed than non-c-commanding QP in on-line processing. Their eye movement data showed that some morphological features of non-c-commanding QP could potentially gain access to the irrelevant anaphor. Dillon et al. (2013) demonstrated that c-command had evident effects on the early comprehension of the anaphoric dependency under the interference paradigm, suggesting that anaphors are only licensed by the c-commanding antecedent. Under the NPI analysis of Vasishth et al. (2008) and Parker & Phillips (2016), however, non-c-commanding negative quantifier *no* XP can partially license the structurally incompatible NPI *ever* in the intrusive environment (i.e. c-command relation). These findings cast serious problems on how human processors access relational information (i.e. X c-commands Y), questioning whether c-command is the “soft and malleable” constraint or the “hard and unintrusive” constraint.

In the processing literature, this reading behavior has been called “linguistic illusion” in which human readers are sometimes inclined to partially accept a violation of the syntactic scope (i.e. c-command relation (Parker & Phillips, 2016). For instance, the ill-formed example (6) containing NPI *ever* is partially acceptable because human subjects can retrieve the semantically plausible licensor *no* in the structurally inappropriate position (Xiang et al., 2009).

(6) ??The restaurants that *no*$_i$ local newspapers have recommended in
their dining reviews have *ever*$_i$ gone out of business
(7) *The new executive who oversaw the middle *managers*$_i$ apparently
doubted *themselves*$_i$ on most major decisions

Notably, Xiang et al. (2009) found event-related potential (ERP) evidence that the sentence like (6) is not acceptable for patients with autistic disorder contrasting to the linguistic illusion of normal human subjects. They concluded that the damaged pragmatic reasoning of patients did not boost the acceptability of the incompliant NPI licensing. This implies that there is an interaction between the intrusion of the syntactic c-command constraint and the offsetting pragmatic reasoning in the NPI domain.

In the reflexive binding domain, however, the intrusion effects on the c-command constraint were not found. See the example (7) where the linguistic intrusion could possibly be triggered, but it was not. Embedding the plural DP *managers* that potentially matches the agreement condition for the reflexive *themselves* did not reduce the reading time in the eye-tracking experiments (Dillon et al., 2013). Overall, the interaction of the c-command constraint with the intrusion effects during sentence processing is dependent on the types of licensing and the reading capacity of human subjects. In this regard, studying the linguistic illusion can analyze how language processors access the structured representation of relational information in memory.

Enlarging upon these findings of the syntactic processing of relational information, recent computational approaches assessed grammatical knowledges of neural language models using the presented psycholinguistic methodology (Futrell et al., 2019; Lin et al., 2019; Wilcox et al., 2019). Specifically, Marvin & Linzen (2018) used context-free grammar to build artificial language data of the targeted linguistic constructions in question. But their evaluative method did not exclude structurally simple constructions like (8a) and (9a) which do not have a modifier clause.

(8) a. Few/Most students have *ever* seen the bird.
    b. Few/Most students <u>that left the class</u> have *ever* seen the bird
(9) a. The engineer/engineers have doubted *himself*/*themselves*.
    b. The engineer/engineers <u>that fixed the car</u> have doubted *himself*/*themselves*.

As such, structurally shallow constructions cannot identify whether a language model is sensitive to the hierarchical c-command constraint. In structurally complex constructions like (8b) and (9b), however, a language model taking a linearly adjacent DP as an antecedent implies that it trivially favors the linear precedence order over hierarchical generalizations. Accordingly, the inclusion of structurally simple constructions can possibly obscure the assessment of a language model if we focus on the syntactic capacity.

Warstadt et al. (2019) explored the linguistic knowledge of fine-tuned BERT using the whole gamut of NPI licensing contexts investigated in the theoretical literature. Under the 2 × 2 × 2 factorial design experiment (Licensor × Scope × NPI), they accommodated additionally selected 8 types of NPI licensors such as rhetorical questions and *if*-conditionals into five variants of the evaluative method. Correspondingly, Warstadt et al. (2019) concluded that the syntactic knowledge of BERT showed a large variability across the evaluative methods, impeding the reliable interpretation of BERT's understanding of the underlying syntactic operation. In addition, Hu et al., (2020) evaluated the syntactic generalization of architecturally different neural language models under 2 × 2 factorial design. Their analytic evaluation aggregated the model behaviors across the 7 types of syntactic phenomena including filler-gap dependency and cleft constructions.

More importantly, all these methods have some limitations in retrieving the encoded syntactic information. This is because measuring the output probability from the softmax function cannot directly access the hidden state of neural language models. Instead, Jumelet & Hupkes (2018) additionally trained another simple linear deep learning model (i.e. *diagnostic classifier*) to obtain the encoded parse tree information from the hidden state where parameters are updated during the model training. In their probing tasks, linear models correctly classified the labeled NPI licensing domain from the semantically unrelated labels in a supervised manner. But predictions of the diagnostic classifier still require sufficient amounts of labeled data and thereby accuracy of the diagnostic classifier is largely dependent on the model training. In summary, the discussion above emphasizes the importance of employing the targeted linguistic items (i.e. reflexive anaphors and NPIs) and the relevant syntactic constraint (i.e. X c-commands Y) for the systematic assessment of neural language models.

3. Methods

3.1. Data Composition: Reflexives

Our study used a 2 × 2 design: C-command × relevant features. In the reflexive pronoun domain, we manually crafted 75 sentences of the 2 × 2 design defined by the c-command condition and the [Number] feature (300 sentences in total). All the sentences contain a subject relative clause as in Table 1. Slashes indicate regions of a subject relative clause.

Table 1. Example items for Experiment 1

Reflexive anaphor	Example Sentences
not C-commanded [singular]	*The **doctors**$_i$ / that ignored the **patient**/ protected *himself*$_i$
C-commanded [singular]	The **doctor**$_i$ / that ignored the **patients**/ protected *himself*$_i$
not C-commanded [plural]	*The **doctor**$_i$ / that ignored the **patients**/ protected *themselves*$_i$
C-commanded [plural]	The **doctors**$_i$ / that ignored the **patient**/ protected *themselves*$_i$

As shown in Table 1, we probed the comprehension of the reflexive anaphor *himself* and *themselves* with its co-indexed antecedent via the morphological agreement in [Number]. In addition, the reflexive anaphor needs to be in the c-command domain of the structurally licit antecedent that matches the [Number] feature (i.e. [singular] or [plural]). For instance, the singular reflexive *himself* cannot be bound to the antecedent *the patient* in the embedded clause, while the singular DP *the doctor* in the main clause can bind the reflexive pronoun. If the structurally illicit antecedent is retrieved, then the comprehension of the given reflexive is misguided (Dillon et al., 2013). Thus, each sentence pair of the same number feature is a minimal pair in terms of the c-command constraint, and the ill-formedness indicates that the c-command condition is not met.

3.2. Data Composition: NPIs

Likewise, as for NPIs, we manually crafted 75 sentences of the 2 × 2 design specified by the c-command condition and the [Negative] feature (300 sentences in total). All the sentences contain a subject relative clause as in the Table 2. Slashes indicate regions of a subject relative clause.

Table 2. Example items for Experiment 2

Polarity item	Example Sentences
not c-commanded [+ Negative]	***The** police/ that killed **no** wife/ has *ever* hated the city
c-commanded [+ Negative]	**No** police/ that killed **the** wife/ has *ever* hated the city
not c-commanded [- Negative]	**??No** police/ that killed **the** wife/ has *always* hated the city
c-commanded [- Negative]	**The** police/ that killed **no** wife/ has *always* hated the city

As shown in Table 2, the NPI *ever* requires a negative context to be licensed while the polarity-neutral item *always* does not (Vasishth et al., 2008). Furthermore, the NPI *ever* needs to be in the c-command domain of the structurally appropriate licensor *no* that constitutes a negative context. In principle, the deeply embedded negative quantifier *no* in the subject relative clause cannot entirely license the NPI *ever* in the main clause (Xiang et al., 2009). If the irrelevant NPI licensor is retrieved, then the comprehension of the given polarity item could fail. Accordingly, each sentence pair of the same semantic feature, either negative or non-negative, is minimally different under the assumption that the intrusion of the c-command constraint will lead to an ill-formed construction.

3.3. Model

In our experiments, we used *Bert-base-uncased* and *Bert-large-uncased* publicly released by Google AI. Two variants of BERT models share the same pre-training tasks and transfer learning paradigm as described in the section 2.2. Base BERT has 12 layers, 12 attention heads and the 768 hidden embedding size. Large BERT has 24 layers, 16 attention heads and the 1024 hidden embedding size. For total parameters, base BERT has 110M parameters, and large BERT has 336M parameters.

3.4. Measures

Importantly, this study combines the information-theoretic measure of surprisal (Levy, 2008). It has been employed to probe the neural language model performance as a parallel method of measuring word-by-word human reading time in eye-tracking experiments (Futrell et al., 2019; Wilcox et al., 2019). A comparable perplexity measure (i.e. a joint probability of a sequence of words), however, is originally referred to as a way of intrinsically assessing the language models regardless of real-world applications. Accordingly, computing perplexity is less cognitively plausible than surprisal which is known to be strongly correlated with reading time. Thus, we adopted the former computational psycholinguistic assumption of surprisal for the investigation of neural language model behaviour.

In the surprisal hypothesis, the inverse log likelihood of the conditional probability for the current token directly correlates to the processing difficulty which can also be manifested in reading times (Smith & Levy, 2013). In the logarithmic surprisal equation, t_i is the current word and $t_1. \cdots t_{i-1}$ constitute the preceding context (The letter C denotes extra-sentential contexts).

$$Difficulty(t_i) \propto Surprisal(t_i) = -\log_2 P(t_i | t_1 \dots t_{i-1}, C)$$

The surprisal operates as a causal bottleneck where the comprehension difficulty is incorporated into the word probability (Levy, 2008). If the structural representation of the preceding context t_1. ⋯ t_{i-1} is highly constrained, then the predictability of the following word t_i increases (i.e. reduced reading time) and the inverse log probability drops. For instance, the surprisal rate of the current word can be significantly lower if the subcategorization option of the prior context is relatively more likely and thus more predictable (e.g. a verb can be followed by either the DP or the clausal complement CP).

In our experiment, we assume that the surprisal hypothesis is also applicable to the probabilistic language model like BERT as it is pre-trained to predict the most likely word in a given context. We reconstructed the test materials to compute the surprisal rates of reflexive anaphors and NPIs in the [MASK] position (Table 3).

Table 3. Example items for masking procedures

[MASK]	Masked Example Sentences
{*himself/ themselves}	The **doctors** that ignored the **patient** protected [MASK]
{himself/ *themselves}	The **doctor** that ignored the **patients** protected [MASK]
{*ever/ always}	**The** police that killed **no** wife has [MASK] hated the city
{ever/ *always}	**No** police that killed **the** wife has [MASK] hated the city

[MASK] is a positional dummy word that must be filled out by the test items. Each surprisal value is calculated from the softmax activation layer of BERT where the parameters of word vectors were compressed into the probability distribution ranged over [0-1]. As assumed, a higher surprisal rate indicates that the tested condition in question is highly irregular under the syntactic expectation of BERT. Hence, surprisal value informs the extent to which the model has acquired syntactic expectation regarding the sentence structure.

4. Results

BERT surprisal data were analyzed using the linear regression model in the R programming environment, and this statistical model was fitted following the method of

Bodowinter (2019). The surprisal data was first framed into *tibble* configuration from *tidyverse* package (Wickam, 2017) and then *broom* package (Robinson, 2017) is used to obtain statistics and p-values.

4.1. Experiment 1: Reflexives

The mean values of BERT's surprisal to the c-command condition in the reflexive binding domain are shown in Figure 1.

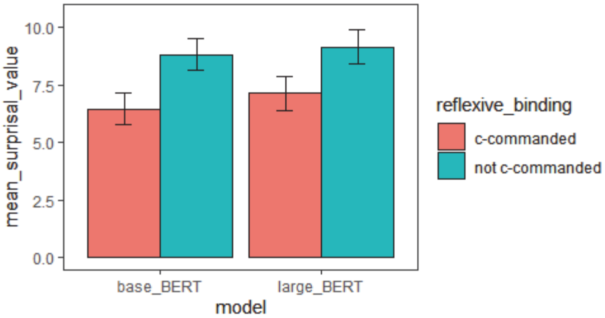

Figure 1. Mean values of BERT's surprisal to the c-command condition in reflexive binding and standard error.

The surprisal analysis of base BERT responses found a positive effect of the structural position of the potential antecedent (*Est.* = −2.35, *SE* = 0.35, *t* = 29.06, *p* < .001). As such, the reflexive pronouns not properly c-commanded by their respective antecedents had a higher average surprisal rate (*mean* = 8.83, *sd* = 2.75) compared to the c-commanded reflexive anaphors (*mean* = 6.48, *sd* = 3.28). The surprisal analysis of large BERT responses also found a main effect of the structural position of the potential antecedent (*Est.* = −2.01, *SE* = 0.38, *t* = 28.89, *p* < .001). The ungrammatical reflexive anaphors with non-c-commanding antecedents had a higher average rate of surprisal (*mean* = 9.15, *sd* = 2.83). In contrast, the grammatical condition yielded relatively lower average surprisal rate of (*mean* = 7.14, *sd* = 3.67).

4.2. Experiment 2: NPIs

The mean values of BERT's surprisal to the c-command condition in NPI licensing are given in figure 2.

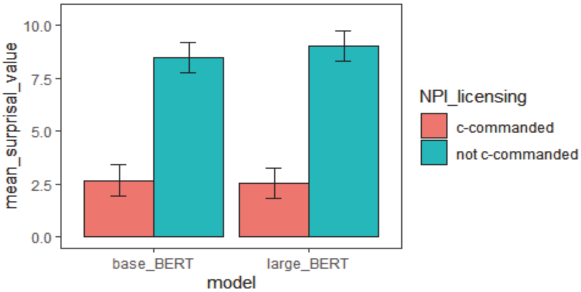

Figure 2. Mean values of BERT's surprisal to the c-command condition in NPI licensing and standard error.

The surprisal analysis of base BERT responses found a significant effect of the syntactic scope of the NPI licensor (*Est.* = −5.81, *SE* = 0.37, *t* = 16.8, *p* < .001). A following observation is that the NPI *ever* had a higher average rate of surprisal (*mean* = 8.49, *sd* = 3.12) when it is not correctly c-commanded by the structurally licit licensor, contrasting to the well-formed string (*mean* = 2.68, *sd* = 3.28). The surprisal analysis of large BERT responses also obtained a main effect of the syntactic scope of the NPI licensor (*Est.* = −6.46, *SE* = 0.37, *t* = 16.9, *p* < .001). The NPI *ever* located outside the c-commanding domain of potential licensors had a higher rate of surprisal (*mean* = 9.02, *sd* = 3.08) while the grammatical condition yielded relatively lower surprisal (*mean* = 2.56, *sd* = 3.36).

4.3. Error Analysis

Assuming that the surprisal rates of BERT generally predict the syntactic effects as shown in the results, we selected some sentences that elicited irregular responses of base BERT (Table 4). The entire examples sentences are listed in appendix and #number indicates the index of examples.

Table 4. Example items for the error analysis

Surprisal	Example Sentences
10.86	***No*** prosecutor that handled **the** felony has *ever* fabricated evidence #102
1.43	*__The__ lecturer that expelled **no** student has *ever* continued the course #147
10.55	The **geologists** that monitored the **worker** supported *themselves* #103
1.39	*The **surgeon** that liked the **teachers** disguised *themselves* #18

Surprisal rates of individual items exhibited a large variability even though the general pattern of results corroborates the claim that the main effect of the c-command condition is statistically significant. Consider the NPI example #102 and #151. The surprisal rate of the c-commanded NPI *ever* is much higher (*Surprisal* = 10.86) than the average surprisal level of the intruded c-command condition. Interestingly, the presence of the ungrammatical NPI *ever* is quite predictable (*Surprisal* = 1.43). A similar pattern is found in the reflexive anaphor example #87 and #34. The surprisal of the c-commanded reflexive pronoun (*Surprisal* = 10.55) is much larger than the ill-formed condition (*Surprisal* = 1.39). All these examples indicate that some other factors intervene the syntactic expectation of BERT.

In our guess, these findings are originated from an inherent problem of language models that are generally constrained by the lexical frequency in training corpora. But our experiment data is unseen during the model training. Furthermore, quantifying surprisal of words mainly integrates the lexical predictability. For instance, negative log-likelihood of the conditional probability such as P(ever|No, prosecutor, the, felony and more) is much lower than the value of P(ever|The, lecturer, no, student and more) regardless of sentence structures. This implies that the acceptability judgments of BERT using surprisal measures are substantially regulated by the statistical frequency in the text corpus. Thus, the direct comparison of individual surprisal rates without a large-scale language experiment is not informative about the encoded syntactic knowledge of BERT.

5. Discussion

The current study assessed whether the BERT language model learned the syntactic scope of c-commanding and non-c-commanding licensors for reflexive anaphors and NPIs. The empirical findings of the first experiment presented the significant statistics showing that two types of BERT are strongly sensitive to the relative position of non-c-commanding antecedents. BERT representations thereby categorically block grammatically inaccessible antecedents that cannot syntactically scope over the cross-clausal reflexive anaphors *himself* and *themselves*. The second experiment also demonstrated reliable statistical evidence consistent with the first experiment. The result of the second experiment on the NPI *ever* suggested that two types of BERT are robustly responsive to the hierarchical restriction of non-c-commanding licensors.

Under the information-theoretic analysis using surprisal measure, we found that either violations of the c-command constraint are highly unacceptable for the BERT language model. The first analysis of these experiments above is that BERT embeddings naturally encode relational information of c-command as visualized in Figures 1 and 2. In these regard, our claim is that the c-command relation is naturally machine-learnable via self-supervision. That is, distributional word embedding models can learn hierarchical rules like a c-command constraint without any explicit supervision of human engineers. But the interesting point is that the grand average surprisal of c-commanded NPI is significantly lower than the c-commanded reflexive pronoun. We argue that this observation is originated from the linguistic sensitivity of BERT. Under our assumption, person match of the anaphoric-dependency is not sufficient for BERT embeddings to clearly comprehend the given anaphoric-dependency in which a grammatical antecedent with the matching [Number] feature c-commands the reflexive pronoun. Contrastingly, the c-commanding NPI licensor *no* strongly licenses the NPI *ever* for the acceptability judgments of BERT. The recent corpus study of de Dios-Flores et al. (2017) also found that a high frequent negative quantifier *no* generates a strong expectation for forthcoming NPIs.

Our second analysis found that surprisal rates of the ungrammatical condition converged around 9.0 when c-commanding licensors are absent for each NPI and reflexive pronoun. In the previous ERP study conducted by Xiang et al. (2009), the NPI illusion elicited P600 effects, which generally indicates a grammatical error as a sign of syntactic anomalies. Inspired by this empirical finding of human brain activities, we propose that surprisal rates around the grand mean surprisal (Surprisal = 9.085) of the tested ungrammatical conditions may generally designate the violation of the c-command relation of given lexical items.

Interestingly, the results of our experiments suggest that acceptability judgments of BERT appear to be drastically diverging from the reading comprehension of human processors encountering reflexive anaphors and NPIs. For human subjects, implanting a non-commanding NPI licensor *no* inside the relative clause triggered a spurious licensing (i.e. violating the c-command requirement for it) of NPI *ever* across many psycholinguistic experiments (Parker & Phillips, 2016). However, BERT clearly distinguished the structural difference between the unlicensed NPI *ever* and the licensed one as in Figure 2. This observation is particularly exceptional since the mean surprisal of unbound reflexive anaphors, which have been reported to be not subject to the interference effects of the hierarchically irrelevant antecedent in the self-paced reading experiments (Dillon et al., 2013), was slightly larger than the bound ones as in Figure 1. If we note that patients with autism disorder did not experience linguistic illusions in NPI licensing (Xiang et al., 2009), One question raised is whether BERT can enter into compensatory semantic and pragmatic reasoning like human processors. To answer this question requires a further investigation on the extent to which BERT fails to replicate sentence processing "errors" which characterize the traits of human language processing.

At the same time, we want to note that in the experiments, architectural components did not affect the surprisal rates relating to c-commandedness. In the context of deep learning, larger parameters and deeper layer-structure generally enhance the model performance in downstream tasks since the bigger model can learn more features and information from data. But it is not clarified whether the model size of BERT has impacts on extracting syntactic information. In our experiments, 340M parameters of the large BERT are not proportionate to decreases in surprisal rates, compared to the base BERT with 110M parameters. This empirical evidence counters the hypothesis that the growth of the size of parameters generally improves on the model performance (Brown et al., 2020). Our results indicate that simply augmenting deep learning architecture does not necessarily boost the BERT understanding of the hierarchical relation.

To summarize, the c-command constraint is rigidly enforced in the BERT language model. The syntactic information is encoded via unsupervised vector representations, whereby the model seeks to predict a masked word in a given sequence of words. The limitation of our study is that the acceptability judgements of BERT cannot be directly compared with human data. This is because our example items are not validated using parallel human language experiments. But the empirical comparison of deep learning data and human data requires a reliable judgements scale that can be cross-validated in a parallel language experiment. In these regard, our study attempted to apply the methods of language experiments for the evaluation of natural language understandings of deep learning models.

6. Conclusion

The present paper integrated the theoretical analysis of syntax and computational method of deep learning. We showed that the structural c-command (i.e. X c-commands

Y) relation is discernable from the contextualized distributional language model like BERT. Under the interference paradigm of psycholinguistic studies, BERT clearly recognized the syntactic scope of c-commanding vs. non-c-commanding licensors with linguistic features. This model capacity suggests that the purely data-driven model with infinitely many decision functions can learn the structural c-command relation from distributional regularities of words.

For the future work, more robustly optimized language models such as RoBERTa and POWER-BERT are available for empirical investigations. These models can elucidate the impacts of the varying training corpus size and different learning patterns on the computational acquisition of grammatical knowledge. Furthermore, language models can be modulated to self-extract the most likely words/phrases from incomplete sentences rather than giving fully completed sentences as our language experiments did. This methodology can examine the semantic and pragmatic inference of the tested language model when it searches for the communicative referent and entailing entity in a given context.

We acknowledge that our language experiments are deterred by the abstract linguistic property of the c-command relation and the black-box system of deep learning models. While disclaiming our data-driven approach is easy, our language experiments, at the very least, provided an empirical ground to corroborate the assumption that syntactic constraints are machine-learnable. In this regard, further studies are anticipated to recruit other empirically tested linguistic phenomena and models - island constraints and BERT, NPI illusion and RoBERTa - for casting fruitful discussions.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). *Language models are few-shot learners*. Computing Research Repository, arXiv: 2005.14165.

Chierchia, G. (2013). *Logic in grammar: Polarity, free choice, and intervention*. Oxford: Oxford University Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of bert's attention. In *Proceedings of the*

2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, 276-286.

Cunnings, I., Patterson, C., & Felser, C. (2015). Structural constraints on pronoun binding and coreference: Evidence from eye movements during reading. *Frontiers in Psychology*, *6*, 840.

de Dios-Flores, I., Muller, H., & Phillips, C. (2017). *Negative polarity illusions: Licensors that don't cause illusions, and blockers that do*. Poster presented at the 30th CUNY conference on human sentence processing, MIT, Cambridge, MA, March 30 – April 1.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 4171-4186.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85-103.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34-48.

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729-743.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 32-42.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4129-4138.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In

Proceedings of the Association for Computational Linguistics, 1725–1744.

Jumelet, J., & Hupkes, D. (2018). Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 222–231

Kim, K.-S. (2010). Is binding possible without c-commanding? *The Journal of Studies in Language, 25*(4), 675-696.

Klima, E. S. (1964). Negation in English. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language* (pp. 246-323). New Jersey: Prentice-Hall.

Kush, D., Lidz, J., & Phillips, C. (2015). Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, *82*, 18-40.

Ladusaw, W. A. (1979). *Negative polarity items as inherent scope relations*. Unpublished doctoral dissertation, University of Texas at Austin.

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

Lin, Y., Tan, Y. C., & Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241-253

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*(1), 99-108.

Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1073-1094.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. Computing Research Repository, arXiv: 1907.11692.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192-1202.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321-339.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry*, *24*(4), 657-720.

Robinson, D., Gomez, M., Demeshev, B., Menne, D., Nutter, B., & Luke, J. (2017). Broom: Convert statistical analysis objects into tidy data frames. *R package version 0.4(2)*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415-433.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685-712.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, 5998-6008.

Wickham, H. (2017). Tidyverse: Easily install and load the 'tidyverse'. *R package version 1(1)*.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211-221.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. London: Routledge.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40-55.

Appendix A. Example items

http://bit.ly/393jv57

Unsub Shin
Graduate Student
Department of Linguistics
Korea University
Anamro 145, Seongbuk-gu
Seoul 02841, Korea
Phone: +82-2-3290-2170
Email: prab35@korea.ac.kr

Myung-Kwan Park
Professor
Department of English Language
Dongguk University-Seoul
30, Phildong-ro 1-gil, Jung-gu,
Seoul 04620, Korea
Phone: +82-2-2260-3153
Email: parkmk@dongguk.edu

Sanghoun Song
Assistant Professor
Department of Linguistics
Korea University
Anamro 145, Seongbuk-gu
Seoul 02841, Korea
Phone: +82-2-3290-2177
Email: sanghoun@korea.ac.kr

Received on January 31, 2021
Revised version received on March 21, 2021
Accepted on March 30, 2021