Linguistic Research 38(2): 395-424
DOI: 10.17250/khisli.38.2.202106.008

Deep learning can contrast the minimal pairs of syntactic data*

Kwonsik Park · Myung-Kwan Park***· Sanghoun Song*****
(Korea University · Dongguk University · Korea University)

Park, Kwonsik, Myung-Kwan Park, and Sanghoun Song. 2021. Deep learning can contrast the minimal pairs of syntactic data. *Linguistic Research* 38(2): 395-424. The present work aims to assess the feasibility of using deep learning as a useful tool to investigate syntactic phenomena. To this end, the present study concerns three research questions: (i) whether deep learning can detect syntactically inappropriate constructions, (ii) whether deep learning's acceptability judgments are accountable, and (iii) whether deep learning's aspects of acceptability judgments are similar to human judgments. As a proxy for a deep learning language model, this study chooses BERT. The current paper comprises syntactically contrasted pairs of English sentences which come from the three test suites already available. The first one is 196 grammatical –ungrammatical minimal pairs from DeKeyser (2000). The second one is examples in four published syntax textbooks excerpted from Warstadt et al. (2019). The last one is extracted from Sprouse et al. (2013), which collects the examples reported in a theoretical linguistics journal, *Linguistic Inquiry*. The BERT models, base BERT and large BERT, are assessed by judging acceptability of items in the test suites with an evaluation metric, *surprisal*, which is used to measure how 'surprised' a model is when encountering a word in a sequence of words, i.e., a sentence. The results are analyzed in the two frameworks: directionality and repulsion. The results of directionality reveals that the two versions of BERT are overall competent at distinguishing ungrammatical sentences from grammatical ones. The statistical results of both repulsion and directionality also reveal that the two variants of BERT do not differ significantly. Regarding repulsion, correct judgments and incorrect ones are significantly different. Additionally, the repulsion of the first test suite, which is excerpted from the items for testing learners' grammaticality judgments, is higher than the other test suites, which are excerpted from the syntax textbooks and published literature. This study compares BERT's acceptability judgments with magnitude estimation results reported in Sprouse et al. (2013) in order to examine if deep learning's syntactic knowledge is akin to human knowledge. The error analyses on incorrectly judged items reveal that there are some syntactic constructions that the two BERTs have trouble learning, which indicates that BERT's acceptability judgments are distributed not randomly. **(Korea University · Dongguk University)**

Keywords deep learning, BERT, syntactic judgment, minimal pair, contrast

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03031616). The authors would like to thank two anonymous reviewers for their comments.

** First author

*** Co-corresponding authors

1. Introduction

The present study proposes a quantitative method that facilitates more empirical research on syntactic phenomena by examining whether deep learning's acceptability judgments properly contrast syntactically appropriate–inappropriate constructions. Syntacticians generally use minimal pairs to investigate syntactic phenomena to focus on a specific syntactic constraint regarding a single word or construction. While this approach is rooted in traditional methodology of phonology, it is also applicable to syntactic studies in that utilizing minimal pairs enables researchers to explain a certain syntactic phenomenon allowing some elements but excluding others in the same syntactic environment. Using the methodology, this research aims at assessing whether deep learning can discern binary distinctions between two syntactically appropriate–inappropriate sentences in a minimal pair. If a deep learning model can detect inappropriateness of syntactic expressions and the judgments are quite similar to human judgments, then syntacticians can make use of deep learning as a supplementary tool to study syntactic phenomena in a more empirical and quantitative approach.

There have been studies that implement models with deep neural networks and evaluate them with minimal pairs (e.g., Marvin and Linzen 2018), but it seems that little research has reported the feasibility of using deep learning as an "infrastructure for modeling human sentence processing" (Linzen 2018) in a specific and arithmetical way. Additionally, much of the previous research focuses on ascertaining whether deep learning models can learn syntactic information, though not examining which syntactic constructions they are vulnerable to understand and why. Da Costa and Chaves (2020), for instance, evaluates deep learning models' ability to learn filler-gap dependencies and demonstrates some neural networks can learn the syntactic information, but the paper does not offer detailed explanations for the networks' sensitivity to specific syntactic phenomena. Furthermore, to our knowledge, no previous study comprehensively investigates whether and how deep learning's understanding of syntactic pairs is similar to human acceptability judgments.

It is necessary to explain how and why a model can(not) learn certain syntactic phenomenon correctly in order to validate the utility of deep learning as a reliable tool to research on syntax. As is well-known, performances of deep learning models are not easy to assess due to the *black box* problem of neural networks (Alain and Bengio 2016; and many others). Nonetheless, evaluating a model's performance with respect to its inner mechanism still has a significance to linguistic research. Otherwise, linguists (particularly, syntacticians) can hardly provide an interdisciplinary study between theoretical linguistics and deep learning techniques.

With this in mind, we attempt to account for which syntactic constructions deep learning can discern and which it cannot, and why it is inclined to learn specific constructions. The present study also compares deep learning's acceptability judgments to human judgments with reference to the magnitude estimation results reported in Sprouse et al. (2013). Thereby, this study investigates how deep learning's syntactic knowledge represents human acceptability judgments.

To this end, the current work makes use of a deep learning language model, BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2018). We evaluate BERT's syntactic knowledge with the three test suites made up of minimal pairs excerpted from several sources including syntax textbooks and published linguistics literature with manifold phenomena. *Surprisal* is used to evaluate BERT's ability to discern which construction is grammatically acceptable or not to test it adopting the well-established psycholinguistic approach. By modeling the way that syntacticians study syntactic phenomena, i.e., assessing an ability to distinguish well-formed constructions from ill-formed ones, demonstrating the feasibility of linguistic research on syntactically contrasted minimal pairs with deep learning would proffer a new approach to expand syntactic research.

1.1 Representing humans' acceptability with deep learning

We admit that deep learning's judgments do not fully represent humans' acceptability. Nonetheless, there are not a few studies which attempt to measure how closely a deep learning model's representation approximates to humans' acceptability such as SuperGLUE (A Stickier Benchmark for General-Purpose Language Understanding Systems; Wang et al. 2020). Although more academic dialogue on whether (and how) deep learning's representation projects acceptability is needed to be continued, we have the following standpoint in the current research. Measuring acceptability is to observe a sort of language usage, and the usage is accumulated in corpus. Therefore, given that implementation of a language model is aimed to represent language users' language faculty employing the corpus, it is valid to reason that acceptability is to a degree mirrored in the language model. However, it is certainly necessary to conduct more research on the validity.

1.2 Outline

This article is structured as follows. Section 2 illustrates the methods of this research. Section 3 presents the results of assessing BERT's syntactic ability with the three test suites. Section 4 discusses the implications of this study. Section 5 concludes the present study and proposes the future research agenda on syntactic phenomena with deep learning technology.

2. Method

2.1 Model

We choose BERT as a proxy for deep learning technology as it is proven to learn syntactic information in many previous studies (e.g., Goldberg 2019). BERT, trained with 3,300M words from the BookCorpus (800M; Zhu et al. 2015) and English Wikipedia (2,500M), "is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" (Devlin et al. 2018: 1). Humans also process sentences bidirectionally, in particularly when processing ambiguous sentences, e.g., *The man killed the king with the knife* (Carnie 2013: 96). Furthermore, there are constructions that we need to process in a right-to-left direction such as cataphora (e.g., Moulton et al. 2018), right dislocation (e.g., Cecchetto 1999), and clause-final focus position (e.g., Song 2017: 59). In this respect, BERT's language processing mechanism is designed cognitively closer to humans', so this bidirectional approach is more powerful than the traditional left-to-right one. BERT also uses self-attention mechanism (Vaswani et al. 2017), which specializes in capturing contextual information of each word by computing weighted averages of the vectors of each token in a sentence. For example, *runs* in a sentence "The dog runs" is different from *runs* in "The man runs a restaurant", because in the latter example the meaning of *runs* is determined by *the*, *man*, *a*, and *restaurant*, possibly resulting in the meaning similar to *operate*, not *dash*.

Devlin et al. (2018) train BERT with two tasks: (i) predicting a special noise token, [MASK], and (ii) making a next sentence prediction. The first task is chosen to allow BERT to represent intra-sentence information. Motivated by the Cloze task (Taylor 1953), a token is masked in a sentence, and the model predicts the masked token. In doing so, the model computes the probability of each word occurring in accordance with its surrounding tokens, i.e., its context, from both left-to-right and right-to-left directions simultaneously. To understand inter-sentence information, BERT is trained with the second task to capture inter-sentential discourse information between two sentences.

We make use of the two versions of BERT: base BERT and large BERT. They differ in model size: base BERT's architecture comprises 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads, whereas large BERT comprises 24 layers of transformer blocks, 1024 hidden units, and 16 attention heads, resulting in base BERT having 110M parameters and large BERT, 340M. The number of a model parameter is related to how much and how deeply the model learns from the input data. Devlin et al. (2018) report that large BERT outperforms base BERT in all the tasks of the General Language Understanding Evaluation (GLUE), which is a recognized wide-coverage multi-task benchmark used to evaluate NLP models (Wang et al. 2018). We expect that large BERT also outdoes base BERT in our test suites.

2.2 Test suites

We extract three test suites from DeKeyser (2000), Warstadt et al. (2019), and Sprouse et al. (2013) to evaluate BERT in a comprehensive way. The first test suite is excerpted from DeKeyser (2000), which is a revised version of those in Johnson and Newport (1989). This test set is a comprehensive and highly elaborated set of assessing English learners' syntactic knowledge. Categorized into 11 types (27 sub-types), the original test items were composed of 200 sentences including four pre-test items. With the pre-test items excluded, Test Suite I comprises 98 correct–incorrect minimal pairs with binary grammaticality values, grammatical and ungrammatical. It is exemplified in (1).

(1) a. The little boy is [speaking/*speak] to a policeman. (present progressive)
b. Tom is reading [a book/*book] in the bathtub. (determiners)

c. Peter made out the check but didn’t [sign it/*sign]. (pronominalization)
d. What [is Martha/*Martha is] bringing to the party? (wh-questions)
e. George says [them much/*much] too softly. (subcategorization)

Second, Test Suite II is extracted from the Corpus of Linguistic Acceptability (hereafter, CoLA; Warstadt et al. 2019). CoLA comprises 10,657 English sentences labeled with two conditions, grammatical and ungrammatical. They are compiled from published linguistics literature of diverse sources. We use four syntax textbooks among the sources to focus only on syntactic phenomena (see Table 1). As not all the sentences in each source are composed of minimal pairs, we manipulated the original sentences by constructing ungrammatical sentences if corresponding grammatical sentences do not exist in the source. For instance, the sentence **There was he in the garden* in Adger (2003) does not have a corresponding grammatical sentence, so we create the grammatical counterpart *There was a boy in the garden*. Conversely, some grammatical sentences that have no corresponding ungrammatical ones are excluded in the test suite as we believe it is not appropriate to decide which word in a sentence should be made an erroneous expression, which can be influenced by our bias (note that CoLA does not include the intention of writing each sentence).

We employ CoLA because it includes diverse aspects of syntactic phenomena excerpted from published syntax textbooks as well as it contains a large amount of test items. Meanwhile, CoLA is one of the tasks of the GLUE benchmark. The reason that we test CoLA data although it had already been tested in the benchmark is that we use a different evaluation metric, *surprisal* (Levy 2008; Hale 2001), from what was used in GLUE benchmark. The GLUE benchmark follows the evaluation metric for CoLA chosen in Warstadt et al. (2019): Matthews correlation coefficient (MCC; Matthews 1975). The MCC, which is known to resolve the potential of providing misleading information from accuracy and F1 score (Chicco and Jurman 2020), is a method of evaluating the quality of binary classifications. We admit that it is also a good evaluation metric, but to approach the data in a psycholinguistic perspective, it is probably more adequate to use *surprisal* to measure BERT’s syntactic knowledge, considering that “syntactic contrasts result from perceptual processing in our brain” (Song and Oh 2017: 390).

Finally, we create the last test suite from Sprouse et al. (2013). In keeping with the typical practice in the field of machine reading comprehension or natural language processing, we decide to add diverse sources from published literature as well as syntax textbooks. Sprouse et al. (2013) make use of English sentences exemplified in a theoretical journal *Linguistic Inquiry* for comparing between informal and formal acceptability judgments of sentences. They tested 148 pairwise syntactic phenomena (296 sentences) selected from 1743 English sentences published in *Linguistic Inquiry* between 2001 and 2010.

The present paper uses only minimal pairs that are computationally tractable for masking. For example, [*It appears that Monday Juanita was running late*] and [**Monday appears that Juanita was running late*] are a minimal pair, but each sentence is not computationally appropriate for masking because for the machine they are different in two discrete aspects, i.e., the position of *Monday* and the insertion of *it*. In addition, we filtered out the expressions that do not have a sentence structure, e.g., [*his*/**him*] *book*. If a masked word is an out-of-vocabulary word that is not included in the vocabulary of a model, i.e., an unknown token for the model, it is also excluded; e.g., tokens such as *infuriate* and *wail* are unknown tokens for BERT. However, if an out-of-vocabulary word is created on purpose, such as *sleeped* for testing test takers' knowledge of irregular verb forms, it is included. The examples are also excluded which include more than one syntactically contrasted elements, e.g., [*The patient was examined carefully*/**The patient examines carefully*] have two contrasted elements, active/passive voice and past/present tense, so we excluded the examples. Table 1 shows the results of manipulating the original test sets.

Table 1. Three test suites

Test suites	Source	Number of original Set	Number of manipulated Set
Test suite I	DeKeyser (2000)	196	186
Test suite II	Adger (2003)	948	342
	Carnie (2013)	870	208
	Miller (2002)	426	62
	Sportiche et al. (2013)	651	182
	Subtotal	2,895	794
Test suite III	Sprouse et al. (2013)	2,400	1,468
Total		5,490	2,448

2.3 Evaluation metric

We employ *surprisal* as an evaluation metric of BERT's acceptability judgments.

Surprisal, which is one of the traditional measures for computational language models (Kim et al. 2020), is exploited to measure *how surprised* neural networks are when encountering a target word in a sequence of words, i.e., a sentence (e.g., Da Costa and Chaves 2020; Wilcox et al. 2019). *Surprisal* of a word is “defined as the negative log-probability of w_i in its sentential context [...] and extra-sentential context [...],” formularized in (2), where $w_{1...i-1}$ represents an already-seen input sequence and CONTEXT represents extra-sentential context. *Difficulty* in the formula means the cognitive difficulty of processing a word that a parser feels; if *surprisal* of a word in a context increases, processing difficulty of the word that the parser perceives also increases because the word does not appear commonly in the context, and if *surprisal* is close to “0”, the word is less likely to occur in the context, or ungrammatical.

(2) $\textit{difficulty} \propto -\log P(w_i|w_{1...i-1}, \text{CONTEXT})$ (Levi 2008: 5)

Grounded on the notion of *surprisal*, a *surprisal* value that a model assigns to a word corresponds to how rarely the word in a given sentence occurs in the training data, i.e., to “the extent to which that word is expected under the language model’s probability distribution” (Wilcox et al. 2019). Accordingly, we can assess a language model’s syntactic knowledge by calculating acceptability of a grammatical–ungrammatical pair by comparing *surprisal* scores of them: if the *surprisal* score of an ungrammatical expression is more than a grammatical one, we can assume that the model has appropriately learned the syntactic well-formedness of the expression. Adopting the method of measuring *surprisal* in Da costa and Chaves (2020), we extract *surprisal* scores of a target word in a sentence by converting a softmax activation of the target into a negative log-probability. As BERT has a bidirectional approach, the formula, (2), is not applicable to BERT’s masked model because (i) it considers only an already encountered input sequence other than the following one in a given sentence, and (ii) it also reflects extra-sentential factors. Accordingly, we adapt the formula of Da Costa and Chaves (2020) designed to apply *surprisal* calculation to BERT’s masked model as follows:

(3) $\textit{surprisal} \propto -\log P(\textit{softmax}(w))$

Surprisal of a masked token w is determined by the probability of w occurring depending on surrounding words in left and right locations.

We manually mask where we intend to insert a target grammatical or ungrammatical expression, and each alternative word is replaced with the [MASK] token, as exemplified in (4):

(4) a. Last night the old lady [MASK] in her sleep. (replaced with [died/*die])
b. The girl cut [MASK] on a piece of glass. (replaced with [herself/*himself])
c. Kevin called Nancy for a date [MASK]. (replaced with [up])
d. Kevin called Nancy [MASK] for a date. (replaced with [*up])

After measuring a word's negative log-probability of occurring in the masked location, the current work takes it that if the *surprisal* score of the ungrammatical expression (*die) is higher than the grammatical one (died) as in (4a), it indicates that BERT perceives that past tense is more appropriate than present tense in the environment at hand; as the surrounding words, i.e., the context, of the alternative words are the same, it is reasonable to assume that BERT has a sense of syntactic use for past and present tense.

The present work also follows the standpoint on how to appropriately measure the convergence rate of acceptability judgments on syntactic phenomena as proposed by Song and Oh (2017). The authors present four methodological frameworks for representing the acceptability of syntactic contrasts in terms of (i) directionality, (ii) goodness, (iii) repulsion, and (iv) intensity of contrasted items. Suppose that there is a spectrum of acceptability representing mental response as gradience between extremely bad to extremely good (see Song and Oh 2017). From the standard of directionality, correct acceptability judgments on two sentences in a minimal pair are determined by whether a good sentence goes to the good side and a bad sentence goes to the bad side. Goodness refers to whether the position of acceptability of a good sentence is located in the right half of the spectrum. Therefore, to represent the goodness of two minimally differing items, we need to pinpoint an absolute value of "0" in the spectrum. Repulsion is a matter of distance between two contrasted items in the spectrum; the acceptability of two contrasted items is far enough away if their acceptability differs considerably. According to the authors, the concept of repulsion is in line with Lakoff (1977)'s statement that acceptability should not be addressed in terms of a dichotomy, i.e., in an either/or distinction, but in a hierarchical system, i.e., in a gradient spectrum. In a similar vein, Song and Oh (2017) also suggest that repulsion is related to hard vs. soft constraints proposed by Sorace and Keller (2005), i.e., distance of acceptability between two contrasted items in the spectrum represents a degree to which the two items differ in acceptability. Finally, intensity refers to how narrowly concentrated acceptability judgments are; if there exist some variations in the acceptability of an expression among speakers, acceptability judgments of the expression spread across the spectrum.

Based on the framework outlined, *surprisal* scores of a model between good and bad expressions in the current research are taken to represent the directionality and repulsion of contrasted items. *Surprisal* represents directionality, as each *surprisal* score represents which item is more surprising, i.e., more unacceptable, as exemplified (5)-(7). The items (examples in Test Suite I) are predicted by base BERT, and bolded words are masked tokens.

(5) a. Last night the old lady **died** in her sleep. (surprisal: 1.6570824)
b. *Last night the old lady **die** in her sleep. (surprisal: 12.632971)
(6) a. The girl cut **herself** on a piece of glass. (surprisal: 2.1163106)
b. *The girl cut **himself** on a piece of glass. (surprisal: 4.2965403)
(7) a. Kevin called Nancy **up** for a date. (surprisal: 2.9706469)
b. *Kevin called Nancy for a date **up**. (surprisal: 10.528346)

Results of calculating *surprisal* scores of test items in (5)-(7) show that each score of ungrammatical expressions is higher than the corresponding one of grammatical expressions.

With respect to repulsion, the gap between *surprisal* scores of each pairing words corresponds to the gap between acceptability scores of the corresponding two words in the spectrum. As mentioned above, the standard for the repulsion of contrasted items is described with hard vs. soft constraints. Violating a hard constraint would lead to a bigger repulsion between grammatical and ungrammatical expressions, and soft constraints would to smaller repulsion. Consider Keller's (2000: 90-94) experiment with three soft constraints (definiteness, verb class, referentiality) and three hard constraints (subject-auxiliary inversion, number agreement, and resumptive pronouns) on extraction from picture NPs. The author reports that there is a significant difference in acceptability between the two types of constraints (to get more a detailed description of the constraints, see Keller 2000: 86). In a similar vein, the present work also investigates the existence of a hierarchy of constraints in the gradience of acceptability from BERT's perspective. For example, we have BERT predict *surprisal* of (8) and (9), extracted from Keller

(2000: 85-87). (8) represents examples of a soft constraint, which is definiteness, and (9) represents examples of a hard constraints, which is number agreement.

(8) a. Which friend has Thomas painted **a** picture of? (surprisal: 2.1712706)
b. ?Which friend has Thomas painted **the** picture of? (surprisal: 1.2611948)
(9) a. Which friend **has** Sarah painted a picture of? (surprisal: 2.981986)
b. *Which friend **have** Sarah painted a picture of? (surprisal: 9.180173)

The *surprisal* scores of (8a) and (8b) show not a big difference, and even (8b) is more acceptable than (8a) in light of BERT, which means that BERT is not sensitive to the definiteness constraint. On the contrary, the scores of (9a) and (9b) reveal that BERT is highly sensitive to number agreement.

The present paper measures a repulsion of contrasted items with the relative proportion of two *surprisal* values for the items in a minimal pair. We use relative proportion rather than absolute proportion because using absolute values is likely to be misleading. Results in (10) and (11) illustrate why it is not appropriate.

(10) a. It is **likely** that Jean left. (surprisal: 7.7254944)
b. *It is **reluctant** that Jean left. (surprisal: 15.612077)
(11) a. John must **not** have eaten. (surprisal: 0.01847287)
b. *John must **not** do have eaten. (surprisal: 7.476947)

The gap between absolute values of (10) and that between those of (11) are almost the same (both values are approximately 7), but it seems to be irrelevant to judge that the repulsion of the two items in (10) is also almost the same as its counterpart of those in (11); the *surprisal* value of (10b) is two times higher than (10a), whereas that of (11b) is around four hundred times. Therefore, we judge that it is reasonable to calculate repulsion with relative proportion. The relative proportion between (10a) and (10b) is 33.1% : 66.9%, while the relative proportion between (11a) and (11b) is 0.2% : 99.8%. In the current paper a repulsion score is an absolute value of subtracting a proportion of an ungrammatical item from that of a grammatical item; i.e., the repulsion of (10) is |0.331 − 0.669| = 0.338, and that of (11) is |0.002 – 0.998| = 0.996.

While directionality is interpreted in terms of accuracy of judgments, repulsion is interpreted in terms of a hierarchical approach, which is in line with Lakoff's (1977) framework. It also has to do with soft vs. hard constraints in Sorace and Keller's (2005) perspective. In the current paper, we adapt the stance of Lakoff's (1977) idea rather than Sorace and Keller's (2005) because we suppose that it does not seem clear to determine which constraint is soft or hard while the hierarchical approach does not require such a classification.

Furthermore, we compare BERT's judgments with human subjects' using experiment results reported in Sprouse et al. (2013): magnitude estimation results (transformed into z-scores ranging -2 to +2). The authors use three formal judgment tasks: two-alternative forced choice, magnitude estimation, and 7-point Likert scale. We choose magnitude estimation results among the three tasks for two reasons: (i) we do not choose the two-alternative forced choice as it is not compatible with deep learning's probability output values as its output is one or the other, and thus it cannot represent a repulsion of contrasted items, and (ii) we deem magnitude estimation to be more compatible with deep learning's judgment than Likert 7-point scale, in that the former represents a more diverse range of acceptability judgment.

As *surprisal* does not have an absolute value of "0", i.e., we cannot specify the midpoint in the gradience of the grammatical-to-ungrammatical spectrum with a *surprisal* score, our methodology does not represent the goodness of contrasted items. In the current research the intensity of contrasted items is also not addressed because according to Song and Oh (2017), repulsion concerns the degree of variability across speakers in acceptability judgment on variation, but judgment agents are only two in this study: base BERT and large BERT.

3. Results

3.1 Directionality of contrasted items

The accuracies of acceptability judgments by the two versions of BERT are shown in Table 2.

Table 2. Convergence rates of test suites I, II and III

Test suites	Source	Base BERT	Large BERT
Test suite I	DeKeyser (2000)	92.5% (172/186)	94.6% (176/186)

Test suite II	Adger (2003)	94.2% (322/342)	93.6% (320/342)
	Carnie (2013)	89.4% (186/208)	88.5% (184/208)
	Miller (2002)	87.1% (54/62)	87.1% (54/62)
	Sportiche et al. (2013)	90.1% (164/182)	87.9% (160/182)
	Subtotal	91.4% (726/794)	90.4% (718/794)
Test suite III	Sprouse et al. (2013)	82% (1204/1468)	82.7% (1214/1468)
Total		85.9% (2102/2448)	86.1% (2108/2448)

Overall, the two versions of BERT are competent at distinguishing ungrammatical sentences from grammatical ones. Regarding the directionality of contrasted items, base BERT and large BERT show similar accuracies, indicating that the number of parameters does not influence the results. In DeKeyser (2000), the threshold score of proficiency in grammaticality is 90% (including the four pre-test items), the point which was scored by the participants who have arrived in North America before the age of 16, which is generally known as the critical period. In this respect, both variants of BERT outperform the proficient level of L2 learners. The accuracies of the sources in Test Suite II are over or approximately 90%. However, the results of Test Suite III are relatively lower than those of the other two test suites. This might be explained by how thoroughly vetted the sentences are, which is discussed in more details in Section 4.

3.2 Repulsion of contrasted items

As mentioned above, the present paper measures the repulsion of contrasted items with relative proportion. We investigate if repulsion changes depending on (i) the version of BERT and/or (ii) the correctness of test items. Figure 1 shows the distribution of (i) the repulsion of base BERT and large BERT (left) and (ii) that of correct and incorrect predictions by BERT (right).

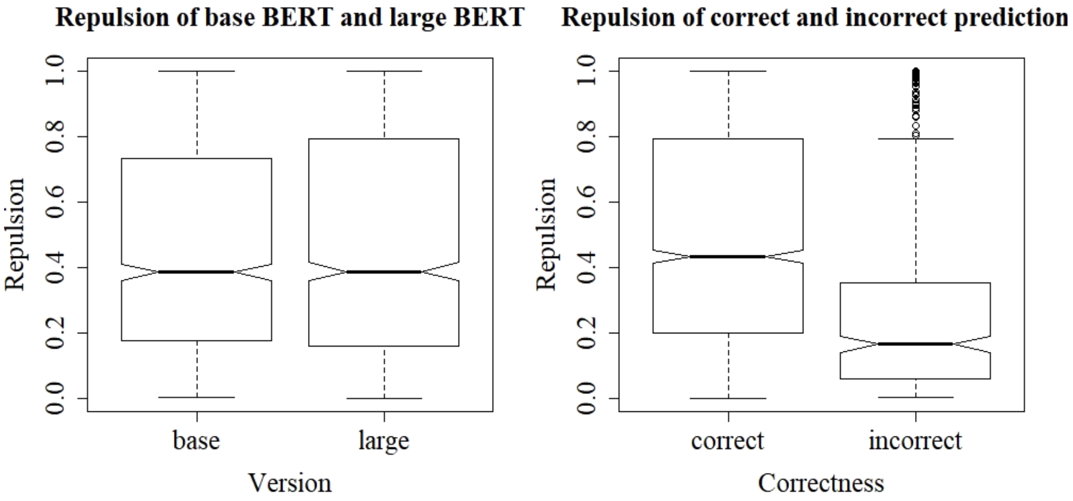

Figure 1. The repulsion of versions of BERT (left) and correctness of predictions (right)

As to the examination of the former, the logistic regression indicates the repulsion of large BERT does not increase or decrease significantly when compared to that of base BERT (b = 0.001267, SE = 0.001232, z = 0.304, OR: 1.001, (95% CI: 0.999, 1.004), $p < 1$ 'ns'). The examination of the latter, on the other hand, indicates that the repulsion of correct decisions by BERT (both variants) increases significantly (though not drastically) when compared to that of incorrect decisions (b = 0.025015, SE = 0.002314, z = 10.811, OR: 1.025, (95% CI: 1.021, 1.03), $p < 0.001$ '***').

The present work also investigates whether the repulsion of judgments on the three test suites differ significantly conducting the two-way ANOVA test. First, the statistical result indicates that all the three sources are significantly different: F (2, 2442) = 14.701, $\eta_p^2 = 0.01$, $p < 0.001$ '***'. Figure 2 represents the distribution of the repulsion of the three test suites.

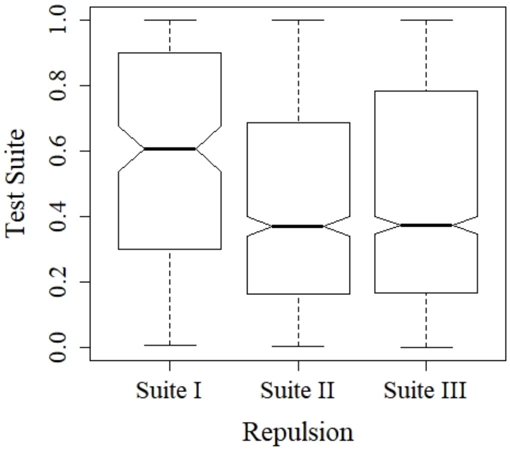

Figure 2. The repulsion of judgments on three test suites

Among the three boxplots in Figure 2, the repulsion of Test Suite I is distributed higher in the plot than those of Test Suite II and Test Suite III. We carry out Tukey's Honestly Significant Difference (HSD) post hoc. It reports that the repulsion of Test Suite I is significantly different from those of Test Suite II (p < 0.001 '***') and Test Suite III (p < 0.001 '***'), but Test Suite II and Test Suite III do not differ significantly in repulsion to each other (p < 1 'ns'). In addition, the repulsion distribution of Test Suite I is significantly higher than those of Test Suite II and Test Suite III as the difference in means indicates: Test Suite I : Test Suite II = 14.395834; Test Suite I : Test Suite III = 12.134907.

Furthermore, the two-way ANOVA test indicates that there is a statistically significant interaction between the effects of correctness and those of types of test suites: F (2, 2442) = 8.936, η_p^2 = 7.27e-03, p < 0.001 '***'. Figure 3 represents the interaction between the effects.

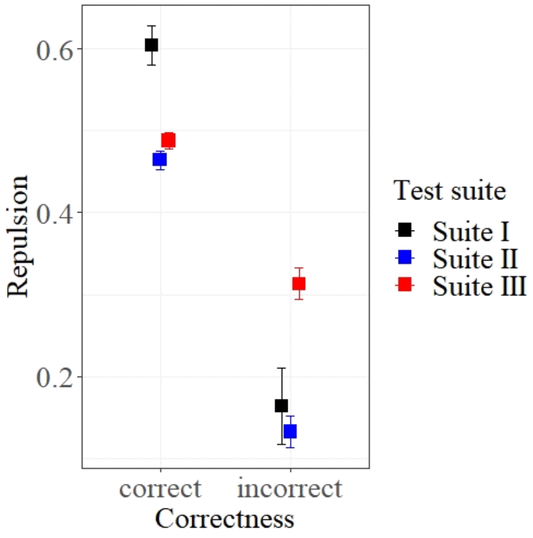

Figure 3. The interaction plot for the effects of correctness and test suites

The repulsion of incorrect predictions is significantly lower than that of correct ones in all the types of test suites (F (1, 2442) = 140.541, η_p^2 = 0.05, p < 0.001 '***'), supporting the logistic regression on correctness mentioned above.

3.3 Error analysis

To analyze which syntactic phenomena BERT fails to distinguish correctly, we list up the ones that BERT cannot, as shown in Table 3. We extract examples of errors (incorrect predictions) only in Test Suite III, which is excerpted from Sprouse et al. (2013), for two reasons. First, using errors in the other test suites can mislead the intended analyses because the errors found in Test Suite I and Test Suite II are not quantitatively sufficient enough to scrutinize, which can result in over-interpretation. Second, BERT's judgment accuracy on Test Suite III is the lowest among the test suites, which means it is worth investigating why BERT shows poor predictions on the items in Test Suite III.

As mentioned before, we address repulsion in terms of a hierarchical approach. For a more detailed perspective, the current paper assumes a standard for interpreting the interaction between accuracy and repulsion, as summarized in Table 1 (We admit that our standard perhaps seems ad-hoc, but we assume this standard to explain the results in a more comprehensive and detailed way). For example, if BERT has intermediate accuracy on a construction and simultaneously it has low repulsion, we can say that BERT is confused about the appropriateness of the construction rather than learns it correctly or incorrectly. BERT's accuracy on the construction is, say, 50%, which is random chance, and repulsion is 0.1. This can be accounted for in the way that BERT is simply confused about the construction without learning the concept of the syntactic phenomena in question. Consider another case where BERT has low accuracy, but repulsion is high. This case would be a situation in which the construction is likely to be highly mis-learned by BERT.

Table 3. The interaction between directionality and repulsion

Accuracy / Repulsion	HIGH	INTERMEDIATE	LOW
HIGH	strongly learned	slightly confused (none in this paper)	strongly mis-learned
INTERMEDIATE	moderately learned	moderately confused	moderately mis-learned (none in this paper)
LOW	weakly learned	highly confused	weakly mis-learned

The incorrectly judged items in Test Suite III are listed up in Table 4 showing the number of errors and the average of repulsion scores

Table 4. Items incorrectly judged by base BERT and large BERT

Idx	Error Type	Number of errors	Repulsion (average)	Example
1	Reflexive pronouns (without any attractor)	30/32 (93.8%)	0.23	Christopher yelled to April to protect [herself/*himself].
2	Negation	30/32 (93.8%)	0.24	Our professor gave [no extensions to any students/*any extensions to no students].
3	*who the hell* (in an interrogative sentence)	30/32 (93.8%)	0.86	[Who the hell/*Who] brought who to the party?
4	Comparative correlatives	30/32 (93.8%)	0.80	[The more/*That much the more] you learn, the more you realize you don't know everything.
5	V-movement	26/28 (92.9%)	0.38	Jack washed the dishes, and Kate [folded the laundry/*the laundry folded].
6	Adjectives	42/48 (87.5%)	0.28	There are leaves [burnt/*green].
7	Multiple quantifiers	28/32 (87.5%)	0.25	[The muffins/*All the muffins] are unlikely to have all been eaten already.
8	Parallel structures	28/32	0.87	We majors in biology are just as smart as

		(87.5%)		you [majors of/*of] mathematics.
9	Voice	88/104 (84.6%)	0.36	The pool [was emptied/*emptied] to find the ring.
10	Verb subcategorization	232/276 (84.1%)	0.32	Mark [sailed to/*sailed] the Caribbean and Cathy flew to the Mediterranean.
11	Adverbial phrases (adjoining to VP)	126/152 (82.9%)	0.42	He envied me [my success after the promotion/*after the promotion my success].
12	*so* serving as complement of a preposition	26/32 (81.3%)	0.49	Megan said that Ben might propose to her, but the actual doing of [it/*so] would take much courage.
13	Omission of complementizer *that*	176/220 (80.0%)	0.21	It seemed at that time [that Michelle had won/* Michelle had won].
14	Determiners	38/64 (59.4%)	0.11	[The cat/*Cat] and dog that were fighting all the time had to be separated.
15	VP ellipsis	38/64 (59.4%)	0.27	steven said he read about a new hybrid car, but I don't know which [car/*car he did].
16	Modality	36/64 (56.3%)	0.52	If Lewis [finishes/*possibly finishes] the report in time, the meeting will be a success.
17	Usage of *likely*	16/32 (50.0%)	0.12	There is likely [to spread a disease/*a disease to spread] around the world.
18	Reflexive pronouns (with an attractor)	58/128 (45.3%)	0.15	John's promise to Susan to take care of [himself/*herself].
19	Adverbial phrases (general)	44/148 (29.7%)	0.29	I only [occasionally bought any books/*bought any books occasionally].
20	VP fronting	0/24 (0.0%)	0.90	Kimberly wanted to give the charity something warm to wear, and give the charity [a bundle of jackets she did/*she did a bundle of jackets].
21	*who the hell* (in a declarative sentence)	0/32 (0.0%)	0.88	I know [who/*who the hell] would buy that book.
22	*there* in interrogative sentences	0/32 (0.0%)	0.88	How many books [were there /*there were] on the table?

There are 22 error types that the two variants of BERT fail to judge correctly in Test Suite III. Given that the correct rate declines drastically after index 13 (Omission of complementizer *that*), we assume that the threshold that we can say BERT understands a syntactic phenomenon is over a correct rate of 80%. In other words, BERT is vulnerable to errors in learning the syntactic phenomena of index number 14 to 22. Figure 4 shows the magnitude estimation results for types of acceptable and unacceptable test items that BERT fails to judge appropriately, as reported in Sprouse et al. (2013).

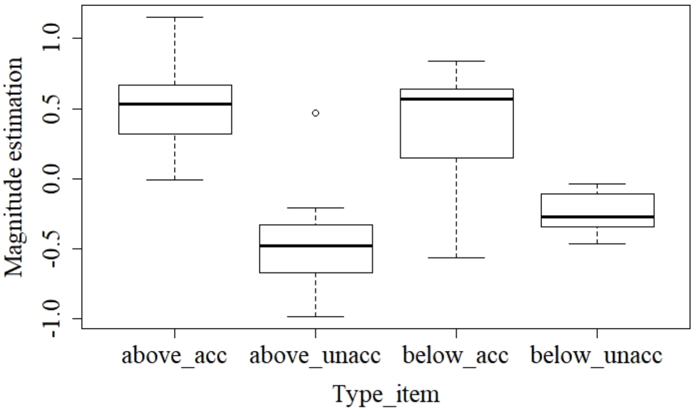

Figure 4. Magnitude estimation of items that BERT failed to predict correctly (above/below = types above/below the threshold, acc/unacc = magnitude estimation of acceptable/unacceptable items)

The more magnitude estimation approximates to 0, the less strong the syntactic contrast of the item is; if the value of an item is 0, it means that even native speakers are confused about whether the item is acceptable or unacceptable. Thus, if the magnitude estimation of items that BERT failed to predict correctly approximates to 0, it would suggest that the confusion with the items of humans is reflected to BERT's low judgment rates. As shown in Figure 3, 'below_acc' and 'below_unacc' seems closer to 0 than 'above_acc' and 'above_unacc'. We carry out a statistical analysis to check if the types of index 1 to 13 in Table 3 are above the threshold and those of index 14 to 22 in Table 3 are below it. We choose the Wilcoxon test because samples are not sufficient enough to carry out the t-test. The statistical result reveals that 'above_acc' and 'below_acc' are not significantly different (W = 65, r = 0.71, p < 1 'ns'). On the other hand, 'above_unacc' and 'below_unacc' differ significantly to each other (W = 24.5, r = 0.71, p < 0.05 '*'). These results indicate that as for at least the error types below the threshold, the ones that BERT failed to learn sufficiently, humans are also confused about the types, relative to the error types above the threshold. This suggest that BERT's judgments are partially similar to humans'.

The following are more detailed analyses for each error type. Error analyses are carried out based on directionality, repulsion and magnitude estimation. We use only magnitude estimation of unacceptable items because acceptable items of below the threshold and those of above the threshold are not significantly different.

3.3.1 Reflexive pronouns

As shown in Table 3, the correct rate of reflexive pronouns is fairly higher when an item does not have any gender attractor (93.8%), which literally attracts an agent to wrong judgments (Bock and Miller 1991), than when an item has an attractor (45.3%). The former indicates BERT has learned binding relation but the latter implies BERT's knowledge of binding is rather heuristic, considering it appears that it takes an antecedent from a candidate which is just closer to the reflexive pronoun in a surface structure (to see more on the heuristic behaviors of deep learning, see McCoy et al. 2019). The following example was taken from Test Suite III, Sprouse et al. (2013); (12a) occurs without any attractor, while (12b) occurs with an attractor.

(12) a. John said/shouted/yelled/screamed/signaled to Sally$_i$ to take care of *her$_i$/herself$_i$/*himself. (Culicover and Jackendoff 2001: 508)
b. John's promise/vow/offer/guarantee/obligation/pledge/oath/commitment to Susan to take care of himself/*herself (Culicover and Jackendoff 2001: 504)

The average repulsion of reflexive pronouns without any attractor is quite low (0.23), which means that for BERT the gender of reflexive pronouns is weakly learned. For reflexive pronouns with and attractor, on the other hand, we can say the considerably low repulsion (0.15) implies BERT is highly confused about the appropriateness of the construction, which is supported by the correct rate which approximates to random chance, i.e., 50% (for more detailed research, deep learning's knowledge of binding rules will be addressed in a forthcoming paper).

The magnitude estimates of unacceptable items of reflexive pronouns with and without attractive are -0.28 and -0.37, respectively, which means humans are confused about the badness of the constructions to a similar degree. We assume that the reason BERT is good at items of reflexive pronouns without any attractor is perhaps BERT judges them correctly just because it selects an antecedent of a reflexive pronoun based on how close a candidate is, rather than it knows the binding rules well.

3.3.2 Adverbial phrases

Word order of adverbial phrases in English is relatively freer than other constituents. This is likely to lead BERT not to comprehend adverbial phrases' word order thoroughly, resulting in BERT being un-strict to the position of them. The correct rate of (general) adverbial phrases (29.7%), which implies BERT does not know how to locate adverbial phrases appropriately in a sentence. However, it seems that BERT has learned one criterion that adverbial phrases are not allowed to be adjoined to VP (Collins and Branigan 1997), which is supported by the correct rate of items assessing adverb phrases adjoining to VP (82.9%). Two relevant examples are sampled in (13); (13a) includes a (general) adverbial phrase, while (13b) includes an adverbial phrase adjoining to VP.

(13) a. *I bought any books only occasionally. (Takano 2003: 521)
b. *Ice cream gives me in the morning brain-freeze. (Johnson 2009: 314)

The repulsion of (general) adverb phrases is also quite low (0.29), but we cannot say BERT is simply confused about the construction because the correct rate does not approximate to random chance. Instead, we can say BERT weakly mis-learns the appropriateness of the construction. The repulsion of adverbial phrases adjoining to VP is 0.42, which means the construction is moderately learned by BERT.

The magnitude estimate of unacceptable items of adverbial phrases (general) over twice larger than that of adverbial phrases adjoining to VP: the former is -0.21 and the latter, -0.48. This means human subjects also feel harder to distinguish the former than the latter as BERT's result indicates.

3.3.3 Usage of *who the hell*

Regarding usage of *who the hell*, Sprouse et al. (2013) design acceptable items comprising only interrogative sentences and unacceptable items consisting of only declarative sentences. For BERT, *who the hell* in both interrogative sentences and declarative ones are acceptable: the correct rate of the former is 93.8% and the latter is 0%. To take an instance, (14a) is an instance of *who the hell* in interrogative clauses, and (14b) is that in declarative clauses.

(14) a. *Who is in love with who the hell? (Dikken and Giannakidou 2002: 56)
b. *I know who the hell would buy that book. (Dikken and Giannakidou 2002: 33)

The repulsion of *who the hell* in interrogative sentences and declarative sentences is 0.86 and 0.88, respectively. This indicates the expression of *who the hell* is highly acceptable to BERT regardless of the two types. The repulsion of the former type indicates BERT strongly learns it. On the other hand, that of the latter type, in particular, implies that BERT strongly mis-learns the construction because it is farthest to random chance.

The magnitude estimates of *who the hell* in interrogative sentences and declarative sentences are -0.84 and -0.04, respectively. This indicates that humans also find it difficult to detect the badness of the latter case, as are BERT's judgments.

3.3.4 VP fronting and *there* in interrogative sentences

The following phenomenon to be discussed is about VP fronting and there in interrogative sentences, as exemplified in (15a-b).

(15) a. *John intended to give the children something nice to eat, and give the children he did a generous handful of candy. (Phillips 2003: 77)
b. *How many books there were on the table? (Hornstein 2007: 410)

VP fronting and *there* in interrogative sentences both have the correct rate of 0%, and the repulsion of them is 0.90 and 0.88, respectively. This indicates BERT strongly mis-learns the constructions. Including *who the hell* in declarative sentences, the three types that has the correct rate of 0% and high repulsion have a common aspect: they are all inappropriate constructions, but the target phrases occur frequently in the corpus (*who the hell* in declarative sentences: 'wh-the-hell' in negated matrix clauses (e.g., Dikken and Giannakidou 2002), VP fronting: 'subject-verb-object', *there* in interrogative sentences: 'wh-subject-verb' in indirect interrogative clauses). Note that we are not arguing this account fully explains the phenomena, but given that deep learning is partially influenced by frequently occurring expression, the account is to some degree explainable.

The magnitude estimates of VP fronting and *there* in interrogative sentences -0.46 and -0.11, respectively. We cannot which estimate score is the exact standard of confusion with a construction, but at least the score of '*there* in interrogative sentences' seems that humans also feel hard to detect its badness.

3.3.5 Other constructions above the threshold accuracy

Error types of comparative correlatives and parallel structure both have not only high accuracy above but also high repulsion, indicating both types are strongly learned by BERT. Error types of negation, v-movement, adjectives, multiple quantifiers, voice, verb subcategorization and omission of complementizer *that* are above the threshold accuracy (above index number 13), but has low repulsion scores, which means they are weakly learned. One the one hand, *so* serving as complement of a preposition is moderately learned by BERT, considering the repulsion (0.49).

Each of error types' magnitude estimate is -0.67 (comparative correlatives), -0.40 (parallel structure), -0.98 (negation), -0.69 (v-movement), -0.21 (adjectives), 0.47 (multiple quantifiers), -0.62 (voice), -0.62 (verb subcategorization), -0.23 (omission of complementizer *that*), and -0.33 (*so* serving as complement of a preposition), respectively. Some estimates such as negation and v-movement indicate that BERT and humans can detect the badness of the constructions, but some estimates such as adjective, in particular, multiple quantifiers reveal that humans are not sensitive to the badness of the constructions, but BERT is not.

3.3.6 Other constructions below the threshold accuracy

The correct rates of determiner, VP ellipsis, and usage of *likely* approximate to random chance and simultaneously the repulsion of them is fairly low. This indicate BERT is confused about the well-formedness of the constructions like in the case of reflexive pronouns with an attractor. Meanwhile, the correct rate of modality is below the threshold accuracy and its repulsion is 0.52, which means BERT moderately confused about the construction.

Each of error types' magnitude estimate is -0.35 (determiner), -0.34 (VP ellipsis),

-0.27 (usage of *likely*), and -0.08 (modality), respectively. Generally, it seems that the estimate scores of these types are close to 0, especially in the case of modality. This indicates humans feel also hard to detect the badness of the constructions, as BERT's accuracy indicates.

4. Discussion

As to the validity of using deep learning as a tool to research on syntactic studies, some would say that it is questionable that deep learning models including learn a language or represent human's sentence processing. This concern is largely because the inner mechanism is not fully accountable. Nonetheless, we would like to quote a proverb, "When you hear hoofbeats, think of horses not zebras." That is, if deep learning models can detect well-formedness of syntactic data good enough and the judgments are made consistently and systemically, then we should say that deep learning is aware of syntax to some degree. The present study has further attempted to articulate why BERT yields such predictions with directionality, repulsion, a comparison with human judgments and error analyses.

The high accuracy of BERT's acceptability judgments on various types of syntactic minimal pairs indicates that deep learning models are of great use as a tool of syntactic research. This would be partially due to the fact that BERT consists of self-attention-based transformers with a bidirectional approach. However, another prediction that the accuracy of large BERT outperforms that of base BERT is not confirmed; the result reveals that the number of parameters of the large variant does not significantly improve syntactic knowledge. Furthermore, given that the distribution of repulsion does not change depending on the two variants of BERT, indicated by the logistic regression, we assume that at least for our test suites, the parameters that are trained by base BERT include linguistic features required for understanding diverse (but not all) syntactic phenomena. With respect to the repulsion distribution of correct and incorrect decisions, although not remarkable, it suggests that the items predicted incorrectly are "hard" for BERT to predict exactly because in the BERT's knowledge they lie lower in the hierarchy of syntactic constraints, where un-strict ones are located.

Considering the repulsion of the test suites, Figure 2 indicates that the repulsion distribution of Test Suite I, which is extracted from DeKeyser (2000)'s stimuli for testing

L2 learners' knowledge of grammaticality, is significantly higher than Test Suite II (excerpted from syntax textbooks) and Test Suite III (excerpted from linguistic literature). This is probably due to the author's intention: testing L2 learners' ability is not the matter of acceptability judgments, but that of grammaticality judgments, i.e., it is reasonable to use test items located higher in the hierarchy of constraints which would be silently learned by learners as well as not be too ambiguous to be judged with confidence.

Regarding the accuracies of each test suite, the performance on Test Suite I and II is notably higher than Test Suite III. As aforementioned, this is partially accounted for by the following reasons. First, in line with Linzen and Oseki (2018), an account for the results is that sentences in sources such as textbooks are more thoroughly vetted by a larger number of linguists for a longer period, leading to weed out controversial cases that do not reach consensus. Ahn et al. (2019) also put forward that a convergence rate is influenced by characteristics of texts, i.e., by the extent to which the texts are empirically tested with enough feedback of linguists. In addition, many examples are made by a linguist's informal judgments, thus possibly resulting in some cases where a paper whose examples do not represent the entire linguists' judgments is published (Cho et al. 2019). In the case of Test Suite I, although not extracted from a textbook but a paper, it is a revised version from a previous acknowledged paper (Johnson and Newport 1989) which stands the test of time. Therefore, the first reason is also applicable to Test Suite I. Furthermore, Test Suite I is designed for assessing learners' grammaticality judgment, thus including only strongly contrasted items, which probably increases the repulsion of items.

Error analyses on incorrectly judged expressions reveal that, except for types whose accuracy approximates random chance (5 out of 22 types), there exist patterns in troublesome learning of BERT. This implies that we can investigate on syntactic phenomena with BERT as it does not just randomly judge them; it shows specific inclination in acceptability judgments. Furthermore, comparing humans' acceptability judgments to those of BERT reveals that judgments of BERT and humans on some phenomena (e.g., adverbial phrases) are similar to each other. This indicates that BERT's prediction reflects humans' cognitive sentence processing to some degree. However, due to the limitation that BERT tends to be more receptive to constructions which occurs frequently in training data, it turns out that the neural language models do not fully represent the human syntactic knowledge as of yet.

In the current paper we do not address all the error types exhaustively except for some notable ones since the aim of this paper is to shed light on the feasibility of using deep learning as a supplementary tool to research on syntactic constructions, demonstrating deep learning can learn syntactic knowledge. Therefore, further research would be to investigate deep learning's inclination to specific syntactic phenomena with more exhaustive methods by using diverse variations of the phenomena such as length, tense, embedded clauses, etc.

In addition, regarding the comparison of BERT's *surprisal* results with humans' magnitude estimation, our measurement has a limitation since it is perhaps more reasonable to use probability of the whole words in a sentence rather than focusing on probability of a word using *surprisal*. Therefore, it is also needed to use an evaluation metric that considers probability of the entire words in a sentence such as the one suggested in Warstadt et al. (2020).

Last, but not least, we should say we do not regard *surprisal* as a perfect scale. *Surprisal* may not be an appropriate evaluation metric in the sense that it has not an absolute value of "0" as mentioned above. Besides, there is no proof for us to claim that *surprisal* of a token fully reflects the intricately associated syntactic (and beyond-syntactic) properties. Yet, we believe at the present stage that *surprisal* is one of the representative methods for evaluating language models' language ability. We will use another method in the future research if a more reliable measure is provided.

5. Conclusion

Not a few studies reveal that syntactic knowledge is manifested in deep learning's knowledge representation. However, the studies do not address which syntactic phenomena is hard for deep learning networks to predict correctly, and few studies compare human and deep learning's judgments of syntactic minimal pairs. In this respect, it has not fully substantiated that deep learning is utilizable for syntacticians to study syntax. Thus, we attempt to demonstrate the feasibility of using deep learning as a supplementary tool to probe into syntactic phenomena.

The experimental results of the present study indicate that syntax research with minimal pairs using BERT is feasible for three reasons. First, the accuracy of BERT's acceptability judgments is fairly high, as shown in the results of directionality of contrasted items. Second, there are specific syntactic phenomena that BERT is vulnerable to discern, as indicated in the error analysis. Third, acceptability judgments of BERT and humans on some syntactic constructions are similar to one another, indicating humans' cognitive responses for some constructions is to some degree represented in BERT's syntactic knowledge.

As mentioned thus far, further research would address various types of syntactic phenomena and features with respect to deep learning models. Additionally, more comprehensive examples are required to be tested in order to verify the deep learning models' syntactic ability. The future study would also test with different languages such as Korean to confirm whether deep learning techniques are cross-linguistically useful for syntactic studies.

References

Adger, David. 2003. *Core syntax: A minimalist approach*. Oxford: Oxford University Press.

Ahn, Hee-don, Hyoung-moon Kim, Jun Zhao, Ji-hee Ha, and Yong-jun Cho. 2019. A comparison of formal and informal acceptability judgments: A case of discrete experimental data. *Korean Language Research* 51: 57-86.

Alain, Guillaume and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644.

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

Carnie, Andrew. 2013. *Syntax: A generative introduction*. West Sussex, UK: John Wiley and Sons.

Cecchetto, Carlo. 1999. A comparative analysis of left and right dislocation in Romance. *Studia Linguistica* 53(1): 40-67.

Chicco, Davide and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(1): 6.

Cho, Yongjoon, Hyoung-Moon Kim, Jun Zhao, Jihee Ha, Rongju Li, and Hee-Don Ahn. 2019. A Comparison of formal and informal Korean acceptability judgments (2). *Studies in Linguistics* 51: 129-160.

Collins, Chris and Phil Branigan. 1997. Quotative inversion. *Natural Language and Linguistic Theory* 15(1): 1-41.

Culicover, Peter W. and Ray Jackendoff. 2001. Control is not movement. *Linguistic Inquiry* 32(3): 493-512.

Da Costa, Jillian K. and Rui P. Chaves. 2020. Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. *Proceedings of the Society for Computation in Linguistics* 3(1): 189-198.

DeKeyser, Robert M. 2000. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition* 22(4): 499-533.

Dikken, Marcel den and Anastasia Giannakidou. 2002. From hell to polarity: “Aggressively non-D-linked” wh-phrases as polarity items. *Linguistic Inquiry* 33(1): 31-61.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

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

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics* 2: 159-166.

Hornstein, Norbert. 2007. A very short note on existential constructions. *Linguistic Inquiry* 38(2): 410-411.

Johnson, Jacqueline S. and Elissa L. Newport. 1989. Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology* 21(1): 60-99.

Johnson, Kyle. 2009. Gapping is not (VP-) ellipsis. *Linguistic Inquiry* 40(2): 289-328.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD Dissertation. University of Edinburgh.

Kim, Euhee, Myung-Kwan Park, and Hye-Jin Seo. 2020. L2ers’ predictions of syntactic structure and reaction times during sentence processing. *Linguistic Research* 37(Special Edition): 189-218.

Lakoff, Robin. 1977. You say what you are: Acceptability and gender-related language. In Sidney Greenbaum (ed.), *Acceptability in language.* 73-86. The Hague, The Netherlands: Mouton Publishers

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

Linzen, Tal. 2018. What can linguistics and deep learning contribute to each other? *arXiv preprint arXiv:1809.04179.*

Linzen, Tal and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: A Journal of General Linguistics* 3(1).

Marvin, Rebecca, and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031.*

Matthews, Brian W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405(2): 442-451.

McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007.*

Miller, Jim. 2002. *An introduction to English syntax*. Edinburgh, UK: Edinburgh University Press.

Moulton, Keir, Queenie Chan, Tanie Cheng, Chung-hye Han, Kyeong-min Kim, and Sophie Nickel-Thompson. 2018. Focus on cataphora: Experiments in context. *Linguistic Inquiry* 49(1): 151-168.

Phillips, Colin. 2003. Linear order and constituency. *Linguistic Inquiry* 34(1): 37-90.

Song, Sanghoun. 2017. *Modeling information structure in a cross-linguistic perspective*. Berlin, Germany: Language Science Press.

Song, Sanghoun and Eunjeong Oh. 2017. What do you mean by contrast in syntax? *Linguistic Research* 34(3): 387-426.

Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11): 1497-1524.

Sportiche, Dominique, Hilda Koopman, and Edward Stabler. 2013. *An introduction to syntactic analysis and theory*. Hoboken, NJ: John Wiley and Sons.

Sprouse, Jon, Carson T. Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134: 219-248.

Takano, Yuji. 2003. How antisymmetric is syntax? *Linguistic Inquiry* 34(3): 516-526.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762.*

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537.*

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461.*

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7: 625-641.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8: 377-392.

Wilcox, Ethan, Roger Levy, and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068.*

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision,* 19-27.

Appendix

1. Contents: Experiment results
2. Website Link: https://bit.ly/3sbjwe8

Kwonsik Park
Graduate Student
Department of Linguistics
Korea University
145 Anam-ro, Seongbuk-gu,
Seoul, 02841, Korea
E-mail: oneiric66@korea.ac.kr

Myung-Kwan Park
Professor
Department of English
Dongguk University
30, 1-gil, Pildong-ro, Chung-gu
Seoul, 04620, Korea
E-mail: parkmk@dgu.edu

Sanghoun Song
Assistant Professor
Department of Linguistics
Korea University
145 Anam-ro, Seongbuk-gu,
Seoul, 02841, Korea
E-mail: sanghoun@korea.ac.kr

Received: 2021. 03. 07.
Revised: 2021. 04. 14.
Accepted: 2021. 05. 17.