https://doi.org/10.24303/lakdoi.2022.30.1.179

Lexical Effects in Island Constraints: A Deep Learning Approach*

Yong-hun Lee
(Chungnam National University)

Lee, Yong-hun. (2022). Lexical effects in island constraints: A deep learning approach. *The Linguistic Association of Korea Journal*, *30*(1), 179-201. This paper examined the lexical effects (a kind of random effect) of each experimental item in English island constraints. For this purpose, this paper adopted (i) the experimental design and dataset in Lee and Park (2018) and (ii) the deep learning model (the $BERT_{LARGE}$ model) in Lee (2021). After the $BERT_{LARGE}$ model was pretrained with the CoLA dataset, the acceptability scores were calculated for all the sentences in the dataset. As in Lee (2021), the acceptability scores in the $BERT_{LARGE}$ model were measured with the numerical values (neither TRUE/FALSE nor Likert scale), which was similar to the magnitude estimation in experimental syntax. After all the acceptability scores were collected, they were normalized into the *z*-scores and statistically analyzed. In this paper, a mixed-effects model was used where both fixed and random effects could be analyzed, but this paper focused on the random effects which were related to the lexicalization of experimental items. Through the analysis, the following was observed: (i) deep learning models could provide some help to make the experimental designs of syntax more sophisticated and fine-grained, (ii) it was possible to examine and control the lexical effects of experimental items with a deep learning model and a mixed-effects model, and (iii) in the case of island sentences, lexical variability was more crucially affected by the factor *Island* than *Location*.

Key Words: island constraints, lexical effects, deep learning, $BERT_{LARGE}$, mixed-effects model

* I wish to thank anonymous reviewers of this journal for their helpful comments and suggestions. All remaining errors are mine.

1. Introduction

In experimental design, it is important to control the effects of other linguistic and non-linguistic factors, while ensuring the effects of the relevant linguistic factors are free to vary. In reality, however, it is difficult to control other factors which were not directly related to the treatment, even though there are some diagnostic methods for the model evaluation, including the randomness of residuals or the normal distribution of the random factors. The model diagnostics can be applied, after a statistical model was constructed (not before the actual experiments).

In the experimental syntax, two random factors are usually adopted in the actual experiments. One is *speaker/individual variation*, and the other was *sentence/lexical items*. Accordingly, many scholars include these two factors in their experiments and analyze them using a statistical analysis such as ANalysis Of VAriance (ANOVA) or a mixed-effects model (Baayen, 2008; Barr et al., 2013; Gries, 2021). However, it is difficult to clearly separate the effects of *speaker/individual variation* and those of was *sentence/lexical items*, because both random effects simultaneously influence the acceptability scores. In addition, the statistical analysis (an ANOVA or a mixed-effects model) is a kind of post-hoc test, which can be conducted after the experiment.

Nowadays, as deep learning technology develops, there are many trials to use the deep learning models in the study of language (Goldberg, 2019; Wang et al., 2019, 2020; Park et al., 2021; Lee, 2021). It was observed that the deep learning models could represent native speaks' intuitions, since the model implicitly contained the intuition of millions of native speakers. As mentioned in Lee (2021), the advantages of using deep learning models in experimental syntax were (i) that the syntactic experiment(s) would be replicable, (ii) that we could use the target and filler sentences as many as possible since computers are not subject to fatigue, and (iii) that we could eliminate the *speaker/individual* variations from the random effects and focus on the *stimulus/item* variations.

This paper used these advantages of deep learning models to propose a method to examine the lexical effects (or lexicalization effects) of experimental items more closely. Because no human informants were necessary in the experiments, we could eliminate one random factor *speaker/individual variation* from the (deep learning) experiment and focus on *sentence/lexical items*.

For this purpose, this paper utilized (i) the factorial design and dataset in Lee and

Park (2018) and (ii) the deep learning model (the $BERT_{LARGE}$ model) in Lee (2021). After the $BERT_{LARGE}$ model was pretrained with the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019) dataset, the acceptability scores were calculated for all the island sentences in Lee and Park (2018). As in Lee (2021), the acceptability scores in this paper were measured with the numeric values (neither binary classification nor Likert scale), which was similar to the magnitude estimation (ME) method in experimental syntax. After all the acceptability scores were collected for the target island sentences, they were normalized into the *z*-scores and statistical analysis was applied to them. In this paper, a mixed-effects model was applied where both fixed and random effects could be analyzed. However, this paper focused on the analysis of random effects, which the lexicalization of experimental items was encoded. Through the analysis, this paper demonstrated how deep learning models could be used for more fined-grained experiments in the study of syntactic phenomena.

This paper is organized as follows. Section 2 introduces previous studies on experimental and deep learning approaches to English island constraints. Section 3 is on the research method, which says about the dataset and how the acceptability scores are measured in the $BERT_{LARGE}$ model. Section 4 enumerates the analysis results of mixed-effects models. In this section, each island constraint was analyzed with a focus on the random effect. Section 5 includes discussions on the contributions of deep learning models to the experimental design of syntax, and Section 6 summarizes this paper.

2. Previous Studies

2.1. Experimental Approaches to Island Constraints

English *wh*-questions are usually constructed by moving *wh*-phrases from their base position to the sentence-initial [Spec, CP] position. In English, the movement is known to be unbounded, but it is also known that *wh*-phrases cannot cross certain kinds of syntactic boundaries which Ross (1967) called these kinds of constructions *islands*. In theoretical syntax, there have been many studies on the island constraints (Chomsky, 1973, 1986; Rizzi, 1990; among many others), where they have tried to explain why *wh*-phrases cannot cross certain kinds of syntactic boundaries such as CP or NP.

Since experimental methods were introduced into syntax in the late 1990s (Bard et al., 1996; Schütze, 1996; Cowart, 1997; Keller, 2000), there have been a lot of experimental

approaches to the syntactic phenomena, including English island constraints (Sprouse et al., 2012; Sprouse and Hornstein, 2013). These studies employed an experimental approach to island constructions and investigated native speakers' intuition on island constraints. Among them, Sprouse et al. (2012) was a milestone for experimental approaches on island constraints. This study developed 2×2 factorial design combinations in (1) and studied four types of island constraints in (2)-(5).[1)]

(1) Factor Combinations in Sprouse et al. (2012)
   a. NON-ISLAND | MATRIX
   b. NON-ISLAND | EMBEDDED
   c. ISLAND | MATRIX
   d. ISLAND | EMBEDDED

(2) *Whether* Islands
   a. *Who* __ thinks that John bought a car?
   b. *What* do you think that John bought __?
   c. *Who* __ wonders whether John bought a car?
   d. *What* do you wonder whether John bought __ ?

(3) Complex NP Islands
   a. *Who* __ claimed that John bought a car?
   b. *What* did you claim that John bought __?
   c. *Who* __ made the claim that John bought a car?
   d. *What* did you make the claim that John bought __?

(4) Subject Islands
   a. *Who* __ thinks the speech interrupted the TV show?
   b. *What* do you think __ interrupted the TV show?
   c. *Who* __ thinks the speech about global warming interrupted the TV show?
   d. *What* do you think the speech about __ interrupted the TV show?

(5) Adjunct Islands
   a. *Who* __ thinks that John left his briefcase at the office?
   b. *What* do you think that John left __ at the office?
   c. *Who* __ laughs if John leaves his briefcase at the office?
   d. *What* do you laugh if John leaves __ at the office?

1) All the (a)~(d) sentences in (2)~(5) were constructed based on the combinations in (1). That is, (2a) had the factor combination in (1a), (2b) had the factor combination in (1b), and so on.

In these sentences, the *wh*-phrases (or fillers) were marked with the italic font and their traces (or gaps) were represented with __.

In the experiment, 173 native speakers participated, and their acceptability scores were measured with a 5-point Likert scale. Then, the acceptability scores were normalized into *z*-scores, and a statistical analysis was conducted. The results were illustrated in Figure 1.

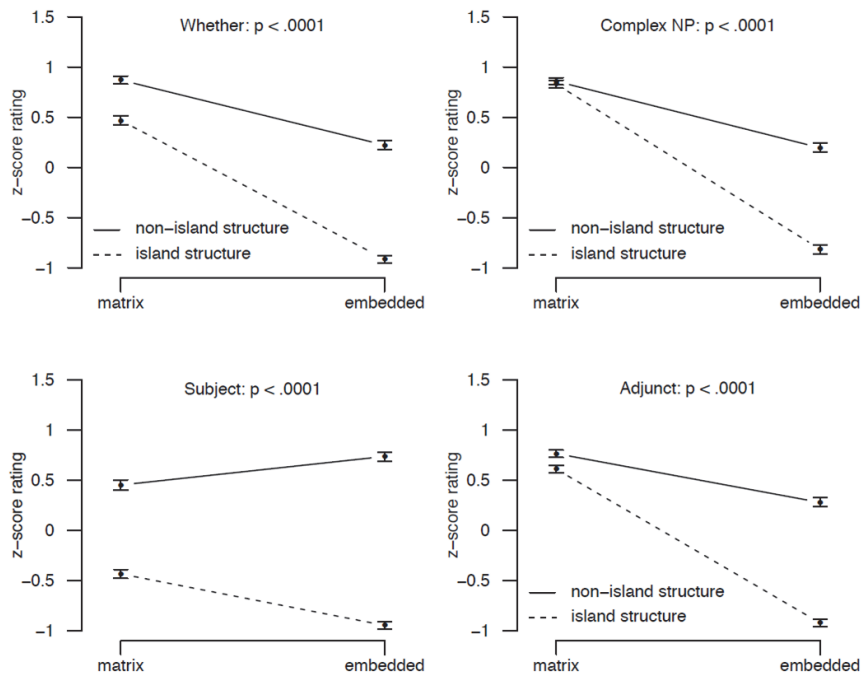

Figure 1. Analysis results of Sprouse et al. (2012)

Figure 1 demonstrated that the acceptability scores of non-island sentences (the solid lines) were much higher than those of island sentences (the dotted lines), and the scores in matrix clauses (the left part) were much higher than those in embedded counterparts (the right part). However, there was one exception (i.e., [non-island, embedded] combination of the Subject island constraints), where the acceptability scores of the [non-island, embedded] were slightly higher than those of [non-island, matrix].

The important findings in their experiment were (i) that the differences between non-island and island sentences in the embedded clauses were much bigger than those in the matrix clauses and (ii) that the differences were statistically significant (p<.0001). This

phenomenon indicated the island effects in English. Thus, their experiment demonstrated that native speakers clearly identified the island constraints.

2.2. Deep Learning Approaches to Island Constraints

Recently, as deep learning technology develops continuously (Goodfellow et al., 2016), there are several approaches to apply the technology in the studies of syntactic acceptability (Goldberg, 2019; Wang et al., 2019, 2020; Park et al., 2021; Lee, 2021). Along with the development of deep learning models, there were also trials to make the dataset for checking the human language faculty. GLUE (Wang et al., 2019) and SuperGLUE (A Stickier Benchmark for General-Purpose Language Understanding Systems; Wang et al., 2020) were developed to measure how closely deep learning models could represent humans' language faculty.

CoLA was the corpus (or the dataset) which collected native speakers' linguistic acceptability to the various types of English sentences (Warstadt et al., 2019). The authors said that "The Corpus of Linguistic Acceptability (CoLA) in its full form consists of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors. The public version provided here contains 9,594 sentences belonging to training and development sets, and excludes 1,063 sentences belonging to a held-out test set."[2] The CoLA dataset became the testbed for the acceptability test of deep learning models.

On the other hand, there have been a few studies which adopted the deep learning technique to analyze various kinds of syntactic phenomena. As for filler-gap dependency or *wh*-movement, three previous studies were noticeable. Wilcox et al. (2018) employed two types of deep learning models and investigated filler-gap dependency and three island effects (*wh*-islands, complex NP, and adjunct). Wilcox et al. (2019a) employed the same models but they extended the scope of the investigation to six islands (*wh*-island, complex NP, subject condition, adjunct, coordination, and sentential subject). Wilcox et al. (2019b) showed that the language model with the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) could be used to learn and identify the filler-gap dependency (or *wh*-movement in Chomsky's theory). In these deep learning models, the acceptabilities were measured not with TRUE/FALSE but with *surprisal* (Levy, 2008).

2) https://nyu-mll.github.io/CoLA/

2.3. Comparisons of Experiments and Deep Learning

Because the island phenomena were independently investigated in both experimental and deep learning approaches, it was difficult to compare the analysis results of the experimental designs and those of deep learning models, since these two types of approaches employed different types of methods for measuring the acceptability scores. Many experimental studies measured the acceptability scores with Likert-scales or ME, whereas deep learning models represented the acceptability scores with TRUE/FALSE (a binary classification) or *surprisal* (Levy, 2008). To overcome this problem, Lee (2021) developed a new deep learning model (the $BERT_{LARGE}$ model), where the acceptability scores were measured with the ME. Then, it was possible to directly compare the analysis results of the experimental designs and those of deep learning models.

Lee (2021) employed the same target sentences in Lee and Park (2018) and calculated the acceptability scores of island sentences.[3] Basically, Lee and Park (2018) followed the factorial design in (1), but four more sets of target sentences were constructed in addition to the sentences (2)~(5).[4] The following sentences illustrated another set of target sentences which were used in Lee and Park (2018).

(6) *Whether* Islands
   a. Who __ thinks that John chased the bus?
   b. What does the police officer think that John chased __?
   c. Who __ wonders whether John chased the bus?
   d. What does the police officer wonder whether John chased __?

(7) Complex NP Islands
   a. Who __ claimed that Mary bought a book?
   b. What did you claim that Mary bought __?
   c. Who __ made the claim that Mary bought a book?
   d. What did you make the claim that Mary bought __?

3) Lee (2021) made use of the same sentences for the target in Lee and Park (2018) but employed different sentences for fillers. In Lee (2021), all the filler sentences came from the CoLA dataset (i) to increase the number of filler sentences and (ii) to use the filler sentences for the evaluation of the $BERT_{LARGE}$ model. For details, see Section 3.5.

4) In Gries (2021), this process was described to make 'concrete token sets'.

(8) Subject Islands
   a. Who __ thinks the sound interrupted the inaugural speech?
   b. What do you think __ interrupted the inaugural speech?
   c. Who __ thinks the sound from the speaker interrupted the inaugural speech?
   d. What do you think the sound from __ interrupted the inaugural speech?

(9) Adjunct Islands
   a. Who __ suspects that the man left the key in the car?
   b. What do you suspect that the man left __ in the car?
   c. Who __ worries if the man leaves the key in the car?
   d. What do you worry if the man leaves __ in the car?

Along with these five sets of target sentences, an identical number of filler sentences were also constructed. As a result, a total of 160 sentences were constructed (4 island types×4 sentence types×5 repetitions×target/filler). In the actual experiments, a total of 100 informants participated who resided in Miami, OH, USA (m=20.340, sd=0.684).[5] The experiments were performed via an online survey using SurveyGizmo.[6]

After all the participants had a warming-up session for the acceptability judgement test, they went into the main task of the experiment. The main task was basically an acceptability judgment task (i.e., intuition test) using the ME method, where all the participants drew different lengths of lines to indicate the naturalness of the given sentence(s). After the experiment was completed, all the acceptability scores for target sentences were collected for each participant. The scores were normalized into the z-scores and statistical analyses were applied to the z-scores. The following plots illustrated the overall results in Lee and Park (2018:447).

As you could observe from the comparison of Figure 1 and Figure 2, the overall tendency in Figure 2 was very similar to those in Figure 1. This implied that the analysis results in Lee and Park (2018) was very similar to those in Sprouse et al. (2012).

In Lee (2021), a deep learning model (i.e., the $\text{BERT}_{\text{LARGE}}$ model) was developed so that the acceptability scores could be measured with the numeric values. The same target sentences in Lee and Park (2018) were used in the experiment and the $\text{BERT}_{\text{LARGE}}$ model measured the acceptability scores of island sentences with values between 0 and 100. The

5) The experiment was approved by the Institutional Review Board (IRB) of the Hannam University (#17-04-01-0201). All subjects involved gave their informed written consent.

6) https://www.surveygizmo.com

scores were normalized into the *z*-scores and statistical analyses were applied to the *z*-scores. The following plots illustrated the overall results in Lee (2021:27).

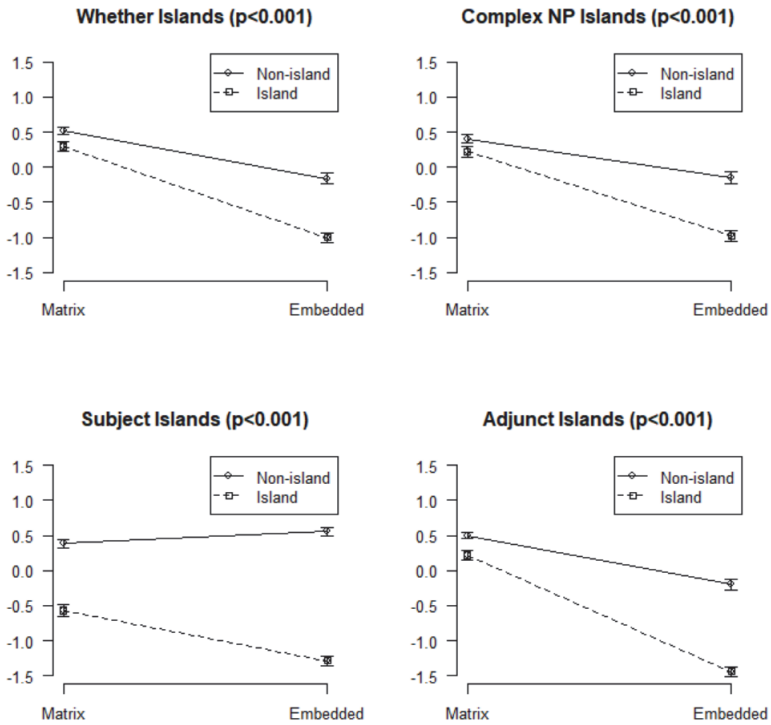

Figure 2. Analysis results of the experimental approach

As you could find, the overall tendency in Figure 3 was similar to those in Figure 1 or Figure 2. This implied that the results of the $BERT_{LARGE}$ model were similar to those of the experimental analysis in Lee and Park (2018). It also implied that the $BERT_{LARGE}$ model correctly reflected the native speakers' intuition on island sentences.[7)]

7) In both experiments, to examine whether or not two linguistic factors (*Island* [island vs. non-island] and *Location* [matrix vs. embedded] in (1)) influenced the acceptability scores of island sentences, Lee and Park (2018) and Lee (2021) conducted statistical analyses. When the normality tests were conducted to the converted acceptability scores (*z*-scores), it was found that most of the datasets did not follow the normal distribution. Therefore, a Generalized Linear Model (GLM) had to be used with a Gaussian distribution (a non-parametric version of ordinary linear regression test) in the statistical analysis of data (Lee, 2016). In both experiments, two linguistic factors (*Island* and *Location*)

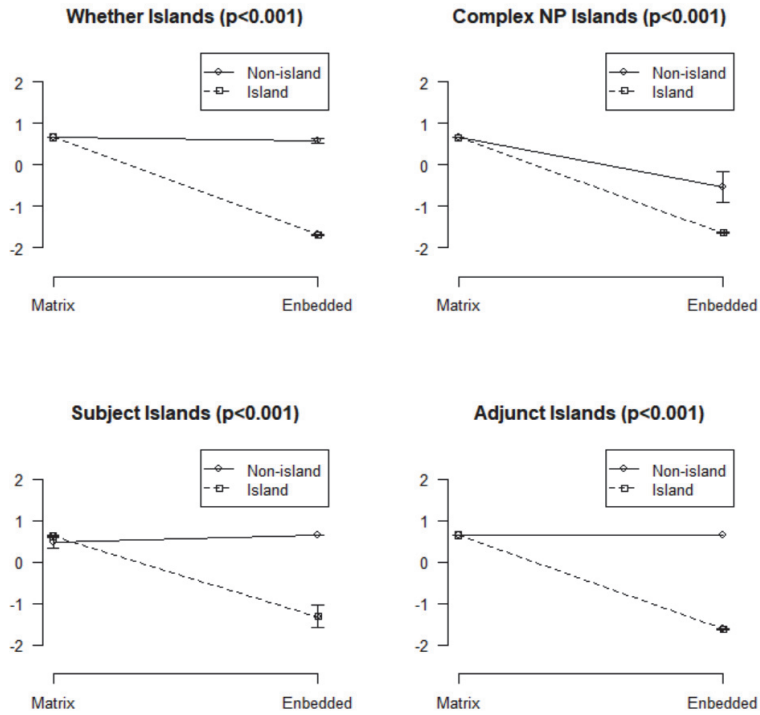

Figure 3. Analysis results of the deep learning approach

3. Research Method

3.1. Dataset

This paper used the same dataset in Lee (2021). As mentioned in Section 2.3, the number of target sentences in Lee and Park (2018) was 80 (4 island types×4 sentence types×5 repetitions). In addition to these target sentences, a total of 400 sentences (80 sentences×5) were randomly extracted from the CoLA dataset. Some sentences could be used

independently influenced the acceptability scores of island sentences (p<.001). In addition, their interaction (*Island*:*Location*) also significantly influenced the acceptability scores of island sentences (p<.001). These results demonstrated that both (American) native speakers and the $BERT_{LARGE}$ model were clearly sensitive to the English island constraints.

as modulus sentences (Sprouse, 2008), and all of the filler sentences were also used in the evaluation of the deep learning model (Section 3.5). After a total of 480 sentences were prepared, they were randomized and they were used as input data to the $BERT_{LARGE}$ model.[8)]

3.2. Deep Learning Model

This paper took the BERT model, because this model was proven to learn syntactic phenomena according to a few previous studies (Goldberg, 2019). In addition, BERT also used a self-attention mechanism (Vaswani et al., 2017), which could capture the contextual information of the sentences while computing weighted averages of the vectors (self-attention) of word tokens in a sentence. According to Devlin et al. (2019), the original English BERT had two versions: the $BERT_{BASE}$ and the $BERT_{LARGE}$. Both models were pre-trained from unlabeled data extracted from the BooksCorpus (Zhu et al., 2015; 800M word tokens) and English Wikipedia corpus (Annamoradnejad and Zoghi, 2020; 2,500M word tokens). Among these two models, this study took the $BERT_{LARGE}$ model, because it had better performance than the $BERT_{BASE}$ model.

After the $BERT_{LARGE}$ model was prepared, it was fine-tuned with the CoLA dataset, since the original BERT was trained with unlabeled data. For this purpose, this paper used the pretrained model in the Hugging Face for consistency.[9)] This was a pretrained $BERT_{LARGE}$ model which was pretrained with the CoLA dataset.

3.3. Procedure

The analysis process in this paper was very similar to that of Lee (2021). The only and important difference was that this study employed mixed- effects models in order to analyze the random effects (lexical effects) of the island sentences.

The procedure proceeded as follows. First, the dataset (a total of 480 sentences) was prepared, and a pretrained $BERT_{LARGE}$ model was downloaded from the website of Hugging Face. Second, the dataset was inserted as an input to the $BERT_{LARGE}$ model, and

8) In the experiments using deep learning models, the randomization process was not necessary since no human beings participate in the experiment. Notwithstanding, the process was applied so that the experiments using the $BERT_{LARGE}$ model were maximally close to the experimental design of syntax. The Latin Square design was not used here, however, because the acceptability scores were not affected by the order of presentation to the computer.

9) https://huggingface.co/yoshitomo-matsubara/bert-large-uncased-cola

the acceptability scores were calculated for each sentence in the dataset, using the algorithm in Section 3.4. Third, after all the fillers (400 sentences) were extracted, the validity of the model was evaluated by the procedure in Section 3.5. Fourth, after all the target sentences (a total of 80 sentences) were extracted, the acceptability scores were converted with *z*-scores. Fifth, statistical analyses (mixed-effects models) were applied to the *z*-scores using R (R Core Team, 2022).[10)]

3.4. Measuring Acceptability Scores

As mentioned in Section 2.3, previous studies measured the acceptability of sentences with a binary classification (TRUE or FALSE; Wang et al., 2019) or with the *surprisal* (Wilcox et al., 2018, 2019a, 2019b).[11)] If the acceptability scores of island sentences could have been measured with these kinds of matrix, however, it was impossible to compare the acceptability scores of experimental design with those of the $BERT_{LARGE}$ model, because the acceptability scores in Lee and Park (2018) were measured with the ME method. Accordingly, a new method was necessary to make the comparisons possible.

The algorithm for measuring the acceptability scores for each English sentence started from the basic architecture of the BERT model.

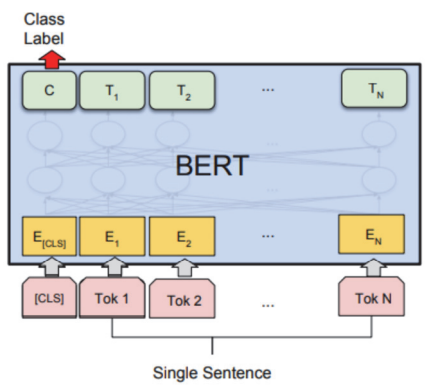

Figure 4. BERT model and CoLA

10) The model was lmer(Score~1+Island*Location+(1+Island*Location|Set)).

11) Basically, *surprisal* (or negative log-conditional probability) tells us how strongly a certain word is expected under the language model's probability distribution (Levy, 2008). If the value was high, it implied that the word occurrence in the given context was a surprise. This implied that the given sentence had more possibility to be unacceptable. There is an (roughly) inverse relationship between *surprisal* and syntactic acceptability.

In the original BERT model, after the model analyzed the input sentence, the model produced a class label, which was TRUE or FALSE.

In Lee (2021), the final output part was revised so that the model could produce two outputs: (i) a class label and (ii) the probability that the given sentence would be TRUE (acceptable). After the probability of TRUE was computed for each sentence in the dataset, the values were normalized with both minimal and maximal acceptability scores in the given dataset.[12] Because the dataset contained clearly acceptable TRUE sentences and clearly unacceptable FALSE sentences, all the values were located between 0 and 1.[13] Then, the values were converted into the acceptability scores, which ranged from 0 to 100.[14]

---

12) Even though clearly acceptable sentences (such as *John was a student*) and clearly acceptable sentences (such as **John were a student*) were included in the dataset, the output probability of TRUE in these sentences might be different depending on the environments of the experiment. The acceptability scores in the deep learning model were calculated against the degree of acceptability in the other sentences. Since (i) not only the filler sentences but also target sentences were included in the dataset and (ii) it was possible to change the targets and fillers in the dataset; the output probability of TRUE in the above two sentences may be different depending on the environments of the experiment. For example, the sentence *John was a student* may have 1.000 of the probability of TRUE in one experiment, but the same sentence may have 0.997 of probability in other experiments. Likewise, the sentence **John were a student* might have 0.000 of the probability of TRUE in one experiment, but the same sentence might have 0.001 of probability in other experiments. The normalization process made the probability of TRUE of *John was a student* 1.000 and that of **John were a student* 0.000, regardless of the experimental environments. The normalization process in the $BERT_{LARGE}$ model was necessary to make the experiment consistent.

13) The clearly acceptable sentences and clearly unacceptable sentences could act modulus sentences to the $BERT_{LARGE}$ model in the sense of Sprouse (2008).

14) Measuring the acceptability scores with *surprisal* and with numeric scores (0~100) had different implications, especially in island sentences. There are roughly two types of accounts which were related to the island constraints in English. The first type of approach is *grammatical accounts* (Chomsky, 1973, 1986, 2000; Lasnik and Saito, 1984; Rizzi, 1990; Szabolcsi and Zwarts, 1993; Tsai, 1994; Reinhart, 1997; Hagstrom, 1998; Truswell, 2007), whose central idea was to explain various types of island constraints under the violation of some grammatical constraints, such as Subjacency Condition (Chomsky, 1973). The second type of approach is *reductionist accounts* or *processing accounts* (Kluender and Kutas, 1993; Kluender, 1998, 2004; Hofmeister and Sag, 2010; Sprouse et al., 2012; Alexopoulou and Keller, 2007), which claimed that the structure-building operations were basically possible also in the island sentences but that the operations wouldn't be carried out in specific circumstances because of some constraints on the resources available to the parsing system (for example, working memory capacity). Strictly speaking, measuring the acceptability scores with *surprisal* is close to the *processing accounts*, while measuring the scores with numeric values (0~100) is close to the *grammatical accounts*.

3.5. Evaluation

After the acceptability scores were measured for each sentence in the dataset, the validity of the obtained scores was evaluated with the fillers (400 sentences). The evaluation proceeded in two steps.

In the first step, the performance of the $BERT_{LARGE}$ model was evaluated with the class labels. Since all the sentences in the CoLA dataset contained the class label (i.e., correct answers), the class labels of all the filler sentences were compared with those in the CoLA dataset. 98.25% of accuracy was obtained in this step.

In the second step, the obtained scores (0~100) were classified into two groups. If the acceptability score was equal to or greater than 50, the sentence had the label TRUE. If not, the sentence had the label FALSE. Then, the labels were compared with the labels of the $BERT_{LARGE}$ model. 97.75% of accuracy was obtained in this step.

From these two steps of evaluation, we expected about 96% of accuracy for the target (island) sentences (0.9825×0.9775=0.9604).

4. Analysis Results

4.1. *Whether* Island Constraint

The following was the results of the mixed-effects model for the *Whether* island constraints. Here, the values for 'Variable 0' were the coefficients that were obtained from the fixed-effects analysis, and those for the others were obtained from the random-effects analysis. Remember that we have five different types of 'lexicalizations' (5 repetitions) in the target sentences.

Table 1. Mixed-effects model analysis in *Whether* island constraint

Variable	(Intercept)	Island	Location	Island:Location
0	**-1.691**	**2.285**	**2.356**	**-2.286**
1	-0.005	0.066	0.006	-0.063
2	-0.010	**-0.276**	0.002	**0.279**
3	0.006	0.057	-0.005	-0.065
4	0.001	0.063	0.004	-0.064
5	0.007	0.058	-0.008	-0.058

This table could be interpreted as follows.

(10) Interpretation of the Model

a. $Score_0$ = (-1.691) + (2.285) × Island + (2.356) × Location + (-2.286) × Island:Location

b. $Score_1$ = (-1.691-0.005) + (2.285+0.066)× Island + (2.356+0.066) × Location + (-2.286-0.063) × Island:Location

c. $Score_2$ = (-1.691-0.010) + (2.285-0.275)× Island + (2.356+0.002) × Location + (-2.286+0.279) × Island:Location

d. $Score_3$ = (-1.691+0.006) + (2.285+0.057)× Island + (2.356-0.005) × Location + (-2.286-0.065) × Island:Location

e. $Score_4$ = (-1.691+0.001) + (2.285+0.063) × Island + (2.356+0.004) × Location + (-2.286) × Island:Location

f. $Score_5$ = (-1.691+0.007) + (2.285+0.058)× Island + (2.356-0.008) × Location + (-2.286-0.058) × Island:Location

That is, the coefficient values in 'Variable 0' became the starting points (or benchmarks), and the values in the other variables indicated the distance from the starting points. Then, the sentence(s) with a 'great absolute value' (i.e., |*value*|) could be the one(s) which behave differently from the others.

In the mixed-effects model for the *Whether* island constraints, the second set of sentences (Variable 2) had the greatest value in *Island*:*Location* (0.279). [15] This implied that this set of sentences was heavily influenced by the lexical effects. The second set of sentences was shown in (6). Also note that the coefficient value was the smallest (0.002) for *Location* in 'Variable 2'. It implied that the value of the *Island*:*Location* column was heavily affected by *Island*, not by *Location*.

4.2. Complex NP Island Constraint

The following was the results of the mixed-effects model for the Complex NP island constraints.

15) Since there were some interactions between *Island* and *Location*, we had to look at the value for *Island*:*Location* first.

Table 2. Mixed-effects model analysis in complex NP island constraint

Variable	(Intercept)	Island	Location	Island:Location
0	**-1.644**	**1.100**	**2.301**	**-1.092**
1	-0.027	0.809	0.014	-0.804
2	-0.026	-0.921	0.038	0.914
3	0.044	-0.109	-0.052	0.109
4	-0.015	-0.972	0.020	0.971
5	0.024	**1.188**	-0.022	**-1.184**

This time, the fifth set of sentences (Variable 5) had the greatest value in *Island*:*Location* (-1.184). This implied that this set of sentences was heavily influenced by the lexical effects, which were shown in (7). Also note that the coefficient value was the greatest (1.188) for *Island*. This implied that the coefficient of the *Island*:*Location* column was heavily affected by *Island*.

4.3. Subject Island Constraint

The following was the results of the mixed-effects model for the Subject island constraints.

Table 3. Mixed-effects model analysis in subject island constraint

Variable	(Intercept)	Island	Location	Island:Location
0	**-1.318**	**1.980**	**1.942**	**-2.127**
1	1.314	**-1.308**	-1.278	**1.455**
2	-0.362	0.361	0.395	-0.210
3	-0.287	0.292	0.322	-0.147
4	-0.353	0.349	0.390	-1.115
5	-0.300	0.296	0.161	0.028

This time, the first set of sentences (Variable 1; the sentences in Sprouse et al. (2013)) had the greatest value in *Island*:*Location* (1.455). This implied that this set of sentences was heavily influenced by the lexical effects. Also note that the coefficient value was the greatest (-1.308) for *Island*. This implied that the value of the *Island*:*Location* column was heavily affected by *Island*. On the other hand, the fifth set of sentences (Variable 5) had

the smallest value in *Island*:*Location* (0.028). This implied that this set of sentences was influenced by the lexical effects at the smallest, which were shown in (8).

4.4. Adjunct Island Constraint

The following was the results of the mixed-effects model for the Adjunct island constraints.

Table 4. Mixed-effects model analysis in adjunct island constraint

Variable	(Intercept)	Island	Location	Island:Location
0	**-1.620**	**2.277**	**2.273**	**-2.279**
1	0.027	-0.027	-0.034	0.029
2	-0.029	0.033	0.033	-0.043
3	0.040	**-0.039**	-0.047	**0.056**
4	-0.019	0.020	0.020	-0.019
5	-0.018	0.014	0.026	-0.021

This time, the third set of sentences (Variable 3) had the greatest value in *Island*:*Location* (0.056). This implied that this set of sentences was heavily influenced by the lexical effects, which was shown in (9). Also note that the coefficient value was the greatest (-0.039) for *Island*. This implied that the value of the *Island*:*Location* column was heavily affected by *Island*.

5. Discussion

5.1. Implications on Experimental Syntax

In the experimental syntax, it is important to control the effects of other linguistic and non-linguistic factors, while the experimental factors are allowed to vary. In reality, it is actually difficult to control all the random factors that were not directly related to the factors under consideration. Even though there are some diagnostic methods for the model evaluation including the randomness of residuals or the normal distribution of the random factors, the model diagnostics can be applied after a statistical model was constructed (not before the actual experiments).

In the experimental syntax, two different kinds of random factors are usually adopted: *speaker/individual variation* and *sentence/lexical items*. Many scholars include these two factors in their experimental design and analyze them with statistical analysis such as ANOVA or a mixed-effects model (Baayen, 2008; Barr et al., 2013; Gries, 2021). However, since the acceptability scores were influenced by both random factors, it is difficult to clearly separate the effects of *speaker/individual variation* and those of was *sentence/ lexical items*. In addition, the statistical analysis such as an ANOVA or a mixed-effects model is a kind of post-hoc test, which can be conducted after the experiments.

In this paper, a method was proposed to control one of the random factors in experimental syntax. The method was to make use of the deep learning models (the $BERT_{LARGE}$ model) and to conduct an experiment before the actual experiments with human participants. Because no human beings participated in the experiments with the deep learning models, it was possible to ignore the random factor *speaker/individual variation* and to focus on the random factor *sentence/lexical items*. After the experiments, statistical analyses such as a mixed-effects model could be applied, and lexical effects of each sentence could numerically be analyzed. As shown in Section 4, the sentence(s) that had the maximal discrepancy could be judged to be the sentence(s) which was/were heavily affected by the lexicalization. If such sentences were found, they could be changed with another set of sentences. It was also possible to take another kind of strategy. Since the deep learning model may have no limitation in the number of target sentences, it was possible to construct enough target sentences (7~8 sets of target sentences). Then, after conducting an experiment with the deep learning model, it would be possible to eliminate a few sets of sentences from the target sentences that were heavily influenced by the lexical variations.

If it was possible to control the effects of lexicalization using the deep learning model, it would be possible to reduce the variations in the acceptability scores and to focus on fixed factors and another random factor *speaker/ individual variation*. That is, the experiments using the deep learning models could contribute to the experimental design of syntax.

5.2. Implications on Island Constraints

The analysis results in Section 4 had interesting implications on island constraints.

First, the more absolute coefficient values the *Island:Location* column had, the more absolute coefficient values the *Island* column had. It implied that the coefficient values in

the *Island*:*Location* column were heavily influenced by the column *Island*. It also implied that the acceptability scores in the island sentences were influenced much more by the factor *Island* than the other factor *Location*.

Second, the coefficient values in the *Island*:*Location* column (or the *Island* column) in *Whether* islands and Adjunct islands were much greater than the values in Complex NP islands and Subject islands. It implied that the latter two types of island constraints were heavily influenced by the lexical effects of the sentences. Note that the CIs of the latter were greater than those of the former.

6. Conclusion

In this paper, a deep learning technique was applied and the acceptability scores were calculated in the deep learning model for each island sentence in English. The dataset came from Lee and Park (2018) and the deep learning model (the $BERT_{LARGE}$ model) came from Lee (2021). After the $BERT_{LARGE}$ model was pretrained with CoLA dataset, the acceptability scores were calculated for each sentence in the island dataset. As mentioned in Lee (2021), the acceptability scores in the $BERT_{LARGE}$ model were measured with the numerical values (0~100), which was similar to the ME method in the experimental syntax. After all the acceptability scores were collected for the target sentences, they were normalized into the *z*-scores and mixed-effects models were applied to them, in order to examine the lexical effects.

Even though this paper utilized mixed-effects models, this study focused on the random effects that were related to the lexicalization of experimental items. Through the analysis, the followings were observed: (i) deep learning models could provide some help to make the experimental designs of syntax more sophisticated and fine-grained, (ii) it was possible to examine and control the lexical effects of experimental items with a deep learning model and a mixed-effects model, and (iii) in case of island sentences, the lexical variability was more crucially affected by the factor *Island* rather than *Location*.

This paper demonstrated how deep learning technology could be actively used in the experimental syntax and how deep learning technology could help experimental design, especially in the control of the lexicaliztion effects. I hope that the combination of deep learning and experimental design can uncover a new perspective in the experimental syntax.

References

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language, 83*(1), 110-160.

Annamoradnejad, I., & Zoghi, G. (2020). ColBERT: Using BERT sentence embedding for humor detection. arXiv preprint arXiv:2004.12765.

Baayen, R. (2008). *Analyzing linguistic data*: *A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language, 72*, 32-68.

Barr, D., Levy, R., Scheepers, C., & Tily, H. 2013. Random effects structure for confirmatory hypothesis testing. *Journal of Memory and Language, 68*, 255-278.

Chomsky, N. (1973). Conditions on transformations. In A. Stephen & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232-286). New York: Holt, Rinehart and Winston.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N. (2000). Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by step*: *Essays on minimalist syntax in honor of Howard Lasnik* (pp. 89-157.). Cambridge, MA: MIT Press.

Cowart, W. (1997). *Experimental syntax*: *Applying objective methods to sentence judgments*. Thousands Oaks, CA: Sage Publications.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. arXiv preprint arXiv: 1901.05287.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Gries, S. (2021). *Statistics for linguistics with R*: *A practical introduction* (3rd edition). Berlin: Mouton

Hagstrom, P. (1998). *Decomposing questions*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Hofmeister, P., & Sag, I. (2010). Cognitive constraints on syntactic islands. *Language,*

86, 366-415.

Keller, F. (2000). *Gradient in grammar*: *Experimental and computational aspects of degrees of grammaticality*. Unpublished doctoral dissertation, University of Edinburgh.

Kluender, R. (1998). On the distinction between strong and weak islands: A processing perspective. *Syntax and Semantics*, *29*, 241-279.

Kluender, R. (2004). Are subject islands subject to a processing account? In V. Chand, A. Kelleher, A. Rodriguez, & B. Schmeiser (Eds.), *Proceedings of the west coast conference on formal linguistics 23* (pp. 475-499). Somerville, MA: Cascadilla Press.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes, 8*, 573-633.

Lasnik, H., & Saito, M. (1984). On the nature of proper government. *Linguistic Inquiry, 15*, 235-289.

Lee, Y., & Park, Y. (2018). English island constraints by natives and Korean non-natives. *The Journal of Studies in Language, 34*(3), 439-455.

Lee, Y. (2016). *Corpus linguistics and statistics using R*. Seoul: Hankuk Publishing Co.

Lee, Y. (2021). English island constraints revisited: Experimental vs. deep learning approach. *English Language and Linguistics, 27*(3), 21-45.

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

Park, K., Park, M., & Song, S. (2021). Deep learning can contrast the minimal pairs of syntactic data. *Linguistic Research, 38*(2), 395-424.

R Core Team. (2022). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Reinhart, T. (1997). Quantifier scope: How labor is divided between QR and choice functions. *Linguistics and Philosophy, 20*, 335-397.

Rizzi, L. (1990). *Relativized minimality*. Cambridge, MA: MIT Press.

Ross, J. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Schütze, C. (1996). *The empirical base of linguistics*: *Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Sprouse, J, Wagers, M., & Phillips, C. (2012). A test of the relation between working memory capacity and syntactic island effects. *Language, 88*, 82-123.

Sprouse, J. 2008. Magnitude estimation and the non-linearity of acceptability judgments. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th west coast*

conference on formal linguistics (pp. 397-403). Somerville, MA: Cascadilla Proceedings Project.

Sprouse, J., & Hornstein, N. (2013). *Experimental syntax and island effects*. Cmabridge, MA: Cambridge University Press.

Szabolcsi, A. 2007. Strong vs. weak islands. In M. Everaert & H. van Riemsdijk (Eds.), *The Blackwell companion to syntax* (pp. 479-531). Oxford: Blackwell.

Truswell, R. (2007). Extraction from adjuncts and the structure of events. *Lingua, 117*, 1355-1377.

Tsai, D. (1994). On nominal islands and LF extraction in Chinese. *Natural Language and Linguistic Theory*, *12*, 121-175.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2020). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537.

Warstadt, A., Singh, A., & Bowman, S. (2019). Neural network acceptability judgments. arXiv preprint arXiv:1805.12471.

Wilcox, E., Levy, R., & Futrell, R. (2019a). What syntactic structures block dependencies in RNN language models? arXiv preprint arXiv:1905.10431.

Wilcox, E., Levy, R., & Futrell, R. (2019b). Hierarchical representation in neural language models: Suppression and recovery of expectations. arXiv preprint arXiv:1906.04068.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? arXiv preprint arXiv:1809.00042.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. arXiv preprint arXiv:1506.06724.

Yong-hun Lee
Instructor
Department of English Language and Literature
College of Humanities, Chungnam National University
99 Daehak-ro, Gong-dong, Yuseong-gu
Daejeon 34134, Korea
Phone: +82-42-821-5331
Email: yleeuiuc@hanmail.net

Received on February 10, 2022
Revised version received on March 21, 2022
Accepted on March 31, 2022