The Journal of Studies in Language

The Journal of Studies in Language 37(4), 507-524 (2022)

Complex NP Island Constraints in English: Experimental vs. Deep Learning Approach

Lee, Yong-hun
Chungnam National University

ABSTRACT

The Journal of Studies in Language **37.4, 507-524**. This paper examined two types of complex NP island constraints (Appositives and Relatives) in English, with an experimental approach and a deep learning approach. In the experimental approach, this paper followed the design in Lee and Park (2018). A total of 120 sentences were employed in the experiment: 40 sentences for the target and 80 sentences for the fillers. For the deep learning approach, this paper utilized the $BERT_{LARGE}$ model that was developed in Lee (2021). The dataset was composed of 240 sentences: 40 sentences for the target and 200 sentences for the fillers. These 240 sentences were used as an input dataset to the $BERT_{LARGE}$ model, and the acceptability scores were calculated for each sentence. After the acceptability scores were obtained for all the target sentences in two different types of approaches, they were normalized into the *z*-scores and statistical analyses were applied to them. Through the analysis, the followings were observed: (i) both the experimental approach and the $BERT_{LARGE}$ model correctly identified two complex NP island constraints in English, (ii) two factors (*Island* and *Location*) and their interaction (*Island*:*Location*) affected the acceptability scores of island sentences, and (iii) two approaches made different predictions on the DD scores of the two complex NP island constraints. **(Chungnam National University)**

Keywords: complex NP island constraints, appositives, relatives, experimental approach, deep learning approach

https://doi.org/10.18627/jslg.37.4.202202.507

pISSN : 1225-4770
eISSN : 2671-6151

Received: January 03, 2022
Revised: February 04, 2022
Accepted: February 12, 2022

1. Introduction

It is known that *wh*-movements are relatively free in English and that they can occur across certain syntactic boundaries, such as TP or CP. In some syntactic environments, however, these *wh*-movements are not available, and Ross (1967) calls them *islands*. There are several different types of islands observed in English and there have been a lot of studies on the islands constraints. The explanation of the island constraints was *grammatical*, *processing-based*, or *experimental*. In the grammatical approaches, the unacceptability of island sentences was accounted for by violation of certain grammatical rules. In the processing-based approach,

the unacceptability was explained by some constraints on the memory resources available to process the sentences. In the experimental approaches, scholars adopted experimental design to measure the native speakers' intuitions and analyzed them with statistical methods.

As deep learning technology is introduced and develops continuously (Goodfellow et al., 2016), there have been several trials to apply the deep learning technology in the studies of acceptability/grammaticality tests (Goldberg, 2019; Wang et al., 2019, 2020; Park et al., 2021; Lee, 2021). Goldberg (2019) demonstrated that recently developed deep learning techniques could be used in the study of syntactic phenomena. Wilcox et al. (2019b) showed that the language model (LM) with the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) could learn and identify not only the filler-gap dependency (*wh*-movement in Chomsky's theory) but also long-distance unbound dependencies. Along with these studies, Wilcox et al. (2018) and Wilcox et al. (2019a) demonstrated that the deep learning model with recurrent neural networks (RNN) was clearly sensitive to English island constraints, which made possible the study of island constraints with deep learning models.

Although there are many different sorts of island constraints in English, this paper focused on two types of complex NP island constraints in English: Appositives and Relatives. Though there are many studies on island constraints, only a few studies compare and examine these two complex NP islands. This paper examined these two kinds of island constraints with two different sorts of approaches: an experimental approach and a deep learning approach. As for the experimental approach, this paper followed the basic experimental design in Lee and Park (2018), and the acceptability scores were measured for native speakers. As for the deep learning approach, this paper used the $\text{BERT}_{\text{LARGE}}$ model that was developed in Lee (2021). The $\text{BERT}_{\text{LARGE}}$ model in Lee (2021) was trained with the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019b) dataset, which was a part of the General Language Understanding Evaluation (GLUE; Wang et al., 2019, 2020). After the acceptability scores were collected from the two different approaches, they were statistically analyzed and compared.

This paper has the following organization. Section 2 introduces previous studies on island constraints and deep learning, which covers previous approaches to island constraints, experimental studies on island constraints, and deep learning analysis to syntactic acceptability. Section 3 is on the research method. This section mentions the dataset for the experimental and the deep learning approach and how the acceptability scores are measured by the $\text{BERT}_{\text{LARGE}}$ model in Lee (2021). Section 4 is on the analysis results. In this section, the statistical analyses are applied to the acceptability scores which are obtained by experiment and the $\text{BERT}_{\text{LARGE}}$ model. Section 5 includes discussions on the differences between experimental data and deep learning data and on the implications on experimental design. Section 6 summarizes this paper.

2. Previous Studies

2.1 Previous Approaches to Island Constraints

There are roughly three different types of accounts which have been provided for 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). The central idea of this

camp was to unify various types of island constraints under a set of generalizations, and the scholars proposed a number of (grammatical) constraints, such as the Subjacency Condition (Chomsky, 1973). This condition said that a *wh*-phrase was not able to cross two or more bounding nodes in one step, and the bounding nodes were noun phrases (NP) and sentences (S). In other words, an island could be said to be a syntactic domain which contained two or more bounding nodes. In Chomsky (1986), the bounding nodes were noun phrases (NP) and inflectional phrase (IP).

The second type of approach is so-called *reductionist accounts* or *processing accounts* (Kluender and Kutas, 1993; Kluender, 1998, 2004; Hofmeister and Sag, 2010; Sprouse et al., 2012; Alexopoulou and Keller, 2007). Scholars in this camp claimed that the structure-building operations were basically possible but that the operations weren't carried out in specific circumstances because of some constraints on the resources available to the parsing system. Specifically, they mentioned that the processing costs of building long-distance dependencies might interact with the processing of island structures and that this interaction caused the burdens in the parsing system. If the burdens exceeded resource capacity (i.e., working memory capacity), people judged the sentences ungrammatical.

The third type of approach is slightly different from the above two approaches. Scholars in this camp employed the experimental methods to syntax (Bard et al., 1996; Schütze, 1996; Cowart, 1997; Keller, 2000; Sprouse, 2008) and measured the native/non-native speakers' intuitions with carefully-designed syntactic experiments and analyzed the corrected data with statistical methods. There have also been many experimental approaches to English island constraints by native speakers (Sprouse et al., 2012; Sprouse and Hornstein, 2013) and non-native speakers (Park and Lee, 2018; Lee and Park, 2018), and these studies used experimental approaches to island constructions and examined native/non-native speakers' intuition on island sentences.

2.2 Experimental Approaches to Island Constraints

Since Ross's identification of island constraints in English (Ross, 1967), there have been a lot of studies on the island constraints in English and other languages.[1] Among the studies, the experimental studies have applied the experimental design and have tried to examine the native/non-native speakers' intuition on island sentences. Among them, Sprouse et al. (2012) was one of the milestones. This study conducted an experiment with island sentences in English and studied native speakers' sensitivity to island constraints. In this study, the scholars adopted 2×2 factor combinations in (1) and tested four different types of island constraints (*Whether*, Complex NP, Subject, and Adjunct) in English, which were listed in (2)-(5).

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

1) Ross (1967) identified seven types of island constraints in English: complex NP, adjunct islands, *wh*-islands, subject islands, left branch islands, coordinate structure islands, and non-bridge islands.

(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?

In these example sentences, the *wh*-phrases were marked with the italic font and the traces with __.

In the actual experiment, the scholars collected a total of 173 English native speakers, and the acceptability scores were measured with 5-point Likert scale. Then, the acceptability scores were transformed into *z*-scores and the statistical analysis were conducted. The results were shown in Figure 1.

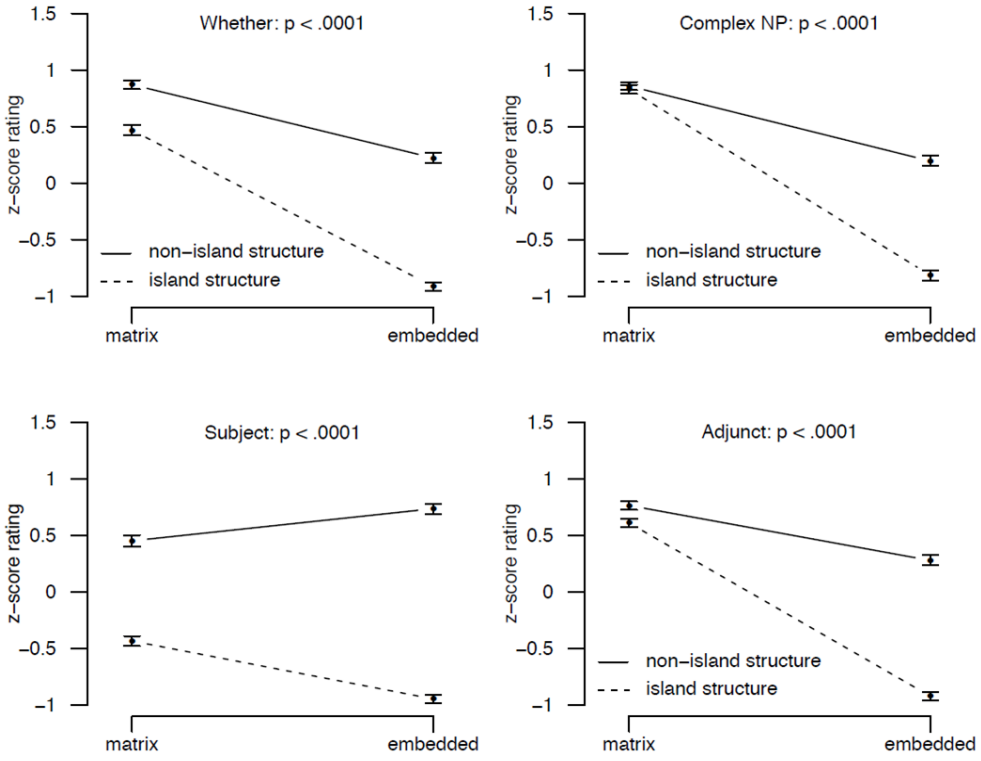

Fig. 1. Analysis Results in Sprouse et al. (2012)

As observed in Figure 1, the *z*-scores (the acceptability scores) in *non-island* sentences were higher than those with *island* sentences, and the scores in *matrix* clauses were much higher than those in *embedded* ones, except in the [non-island, embedded] of the Subject island constraints. An important finding was (i) that the differences (in the acceptability scores) between *non-island* and *island* sentences in the *embedded* clauses were bigger than those in the *matrix* clauses and that (ii) the differences were statistically significant (p<.0001).

Sprouse et al. (2012) claimed that the reason for differences between two types of clauses (*embedded* and *matrix* clauses) was due to the island effects. That is, the island effects were added to the differences between *embedded* and *matrix* clauses, which resulted in the plus values in the differences-to-differences scores (DD scores; Maxwell and Delaney, 2003). These analysis results showed that native speakers in English clearly identified four island constraints.

In Sprouse et al. (2012), however, Appositive sentences in (3) were included in the experiments. Relative clauses, another type of complex NP island constraint, were not included in the experiment, even though the complex NP island constraints included both types of constructions (Carnie, 2021). Szabolcsi (2007) mentioned that both Appositives and Relatives could be strong islands when the CP (within the complex NP) contained a finite tense.

2.3 Deep Learning Approaches to Syntactic Acceptabilities

Since the deep learning technology was introduced and developed recently, there were a few trials to apply the techniques in the study of linguistic phenomena including syntactic acceptability tests. Although the judgments of the deep learning models do not fully reflect native speakers' acceptability scores yet, it is also a fact that the judgments by

deep learning models also represent humans' syntactic acceptability to a considerable extent (Goldberg, 2019; Warstadt et al., 2019b).[2] After this observation, there have been a few trials which attempt to measure the acceptability scores of (English) sentences and to compare them with those of human beings (Goldberg, 2019; Warstadt et al., 2019b; Wang et al., 2019, 2020; Park et al. 2021; Lee, 2021). General-Purpose Language Understanding Systems (GLUE; Wang et al., 2019) and A Stickier Benchmark for General-Purpose Language Understanding Systems (SuperGLUE; Wang et al., 2020) were also developed to provide a standard dataset, with which the scholars could measure how closely a deep learning model's representation of natural languages was approximated to humans' language faculty.

The Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2019b), was the corpus/dataset which collected native speakers' linguistic acceptability judgments to the various types of English sentences. The authors of the CoLA mentioned that "The Corpus of Linguistic Acceptability 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."[3] The CoLA dataset has been the benchmark for testing acceptability scores in the deep learning models.

On the other hand, there were a few studies which applied the deep learning technique to analyze various syntactic phenomena.[4] Marvin and Linzen (2018) employed 350,000 pairs of English sentences and three deep learning models and investigated three linguistics phenomena which were sensitive to the hierarchical structure of sentences: subject-verb agreement, reflexive anaphora, and negative polarity items (NPIs). Warstadt et al. (2019a) used the BERT to show that the BERT was very useful to analyze NPIs. Hu et al. (2020a) utilized the GLUE dataset and six neural language models to study reflexives in English. Likewise, Hu et al. (2020b) used the Brown Laboratory for Linguistic Information Processing 1987-89 Corpus Release 1 (BLLIP; Charniak et al., 2000) and a total of ten language models and examined some linguistic phenomena: subject-verb agreement, two licensing (NPIs and reflexives), garden path effects, gross syntactic expectation (subordination), center embedding, and long-distance dependencies.

As for filler-gap dependency (*wh*-movement in Chomsky's syntactic theories), three different studies were available. Wilcox et al. (2018) adopted two types of deep learning models and studied filler-gap dependency and three island effects (*wh*-islands, complex NP, and adjunct).[5] Wilcox et al. (2019a), on the other hand, adopted the same models but extended the scope of investigations to six islands (*wh*-island, complex NP, subject condition, adjunct, coordination, and sentential subject). Wilcox et al. (2019b) used four different models and to investigate center embedding and

2) Warstadt et al. (2019b), for example, measured the accuracy of the LSTM model for the CoLA dataset and reported that the LSTM model correctly classified 77.2% of the in-domain sentences (into grammatical sentences and ungrammatical sentences) and 73.2% of the out-domain sentences. For the same dataset (the CoLA dataset), the average human judgements were 85.0% for the in-domain sentences and 87.2% for the out-domain sentences, which were slightly higher than the accuracies of the LSTM model. If we admit that the LSTM model is one of the early deep learning models, the discrepancies between humans and deep learning models become much smaller after the introduction of the BERT model, which is based on the Transformer architecture (Vaswani et al., 2017).

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

4) In the studies which were introduced in this section, the specific names of deep learning models were not be provided in most cases. The reasons were (i) that the names contained too many acronyms and abbreviations and (ii) that each study mentioned too many references for their models. It could briefly be said that, however, these models were constructed by the combinations of *n*-gram models, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long-short term memory (LSTM), and so on.

5) The two deep learning models were the Google model (basically, a combination of LSTM and CNN) and the Gulordava model (Gulordava et al., 2018).

various types of filler-gap dependency.

As mentioned in Lee (2021), there were a few reasons why it was necessary to examine the island constraints with deep learning techniques. From the deep learning perspective, it was required to check whether the deep learning model correctly reflected the human beings' intuition to the natural languages or not. From the linguistic point of view, on the other hand, it was also necessary to check if the experimental design correctly reflected the general tendency to the given linguistic phenomena. Because it was impossible to conduct an experiment with all the native speakers in English, all the experimental approaches might have limitations for generalizations. Because it could be said that the deep learning model(s) might include the intuitions of millions or billions of (English) native speakers, the model(s) could be a wonderful testbed of the experimental designs.

3. Research Method

3.1 Dataset

In this paper, two types of dataset were used. First of all, the target sentences in Lee and Park (2018) were adopted without any change. Even though it was not included in the analysis of Lee and Park (2018), the target sentences for Relatives were also constructed and the acceptability scores were also measured with the magnitude estimation at the same time when the experiments in Lee and Park (2018) conducted.[6] The basic format of the target sentences was as follows.

(6) Complex NP Islands (Appositives, equal to (3))
   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 __?

(7) Complex NP Islands (Relatives)
   a. *Who* __ told the boy that he was chasing the dog?
   b. *What* did the man tell the boy that he was chasing __?
   c. *Who* __ saw the boy that was chasing the dog?
   d. *What* did the man see the boy that was chasing __?

For each type, five sets of the target sentences were constructed with lexical variations. As a result, a total of 40 sentences were constructed for the target sentences (4 sentences×5 iterations×2 types).

As for the filler sentences, this study followed the traditions in previous studies and added the filler sentences.[7] In the

6) In the experiments, a total of 100 informants participated who resided in Miami, OH, USA. The mean value (*m*) of their age was 20.340 and the standard deviation (*sd*) was 0.684. 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.

experimental approach, 80 filler sentences were constructed as in Lee and Park (2018). The filler sentences were composed of grammatical and ungrammatical sentences which were not related to the target sentences. Some of the filler sentences could be used as modulus sentences (Sprouse, 2008). In the deep learning approach, 200 filler sentences were constructed, which were different from the fillers in the experimental approach.

The fillers for the deep learning model came from the CoLA dataset, rather than from Lee and Park (2018). In the experimental design, it is usual that the number of fillers has to be 3~5 times the target sentences. Since the number of target sentences was 40, a total of 200 sentences were randomly selected from the CoLA dataset (40 sentences×5). Some of the filler sentences could be used as modulus sentences as in Sprouse (2008), and all of the fillers were also used in the evaluation of the deep learning model (Section 3.5). After 240 sentences were prepared, they were randomized and used as an input data set to the deep learning model.[8] The Latin Square design was not applied for the deep learning model because the acceptability scores for the English sentences did not change by the order of sentences in the presentation to the computer.

3.2 Deep Learning Model

This paper basically used the same deep learning model in Lee (2021): the $\text{BERT}_{\text{LARGE}}$ model pretrained with the CoLA dataset. The BERT model (Devlin et al., 2019) proved to successfully learn syntactic phenomena in natural languages (Goldberg, 2019). BERT employed a self-attention mechanism (Vaswani et al., 2017), which helped to capture the contextual information of each word by calculating weighted averages of the vectors (i.e., self-attention) of each word in a sentence. According to Devlin et al. (2019), the original English BERT had two versions: (i) the $\text{BERT}_{\text{BASE}}$: 12 Encoders with 12 bidirectional self-attention heads, and (ii) the $\text{BERT}_{\text{LARGE}}$: 24 Encoders with 16 bidirectional self-attention heads. These two different models were pre-trained from the unlabeled dataset which was extracted from the BooksCorpus (Zhu et al., 2015; 800M words) and English Wikipedia (Annamoradnejad and Zoghi, 2020; 2,500M words). Among these two models, this study took the $\text{BERT}_{\text{LARGE}}$ model, because it showed better performance than the $\text{BERT}_{\text{BASE}}$ model.

Even though the $\text{BERT}_{\text{LARGE}}$ model was prepared, it was necessary to fine-tune the model with the CoLA dataset again, because the original $\text{BERT}_{\text{LARGE}}$ was trained with the unlabeled dataset of the BooksCorpus and English Wikipedia. For this purpose, Lee (2021) and this paper used the pretrained model in the Hugging Face for consistency.[9] This was a pretrained $\text{BERT}_{\text{LARGE}}$ model which was fine-tuned with the CoLA dataset.

3.3 Procedure

The procedures for measuring acceptability scores were basically identical to those in Lee (2021), and they were as follows. First, the dataset was prepared which contained a total of 240 sentences. Second, a pretrained $\text{BERT}_{\text{LARGE}}$ model was downloaded from the Hugging Face. Third, the prepared dataset was used as an input to the $\text{BERT}_{\text{LARGE}}$

7) Actually, it is possible to make the dataset with only the target sentences for the deep learning model.
8) Actually, the randomization process was not necessary since no human beings participated in the experiment, but the randomization process was adopted here to make the deep learning model maximally close to the experimental design.
9) https://huggingface.co/yoshitomo-matsubara/bert-large-uncased-cola

model and the acceptability scores were computed for each sentence in the dataset, along with the algorithm in Section 3.4. Fourth, after the filler sentences (200 sentences) were extracted, the validity of the model was evaluated with the algorithm in Section 3.5. Fifth, after the target sentences (40 sentences) were extracted, the acceptability scores were normalized with *z*-scores. Finally, statistical analyses were applied to the normalized z-scores using R (R Core Team 2021).

3.4 Measuring Acceptability Scores

The detailed explanations for measuring acceptability scores in the $BERT_{LARGE}$ model were included in Lee (2021). This section included only the basic ideas. Some previous studies in deep learning (such as Wang et al. (2019)) measured the acceptability scores of English sentences with acceptable (TRUE) or unacceptable (FALSE) or with the *surprisal* (Wilcox et al, 2018, 2019a, 2019b).[10] Although these kinds of measurements could reflect acceptability scores to a certain degree, it was difficult to directly compare the acceptability scores of the deep learning algorithm with the results of experimental design in Lee and Park (2018), since the acceptability scores in the experimental design were measured with the scores between 0 and 100. Accordingly, a new method was necessary for measuring acceptability scores in the $BERT_{LARGE}$ model.

The algorithm for measuring acceptability scores in Lee (2021) and this paper started from the basic architecture of the BERT model (Figure 2). According to this figure, after the BERT model processed the input sentence, the model produced a class label (i.e., [CLS]), which was either TRUE or FALSE.

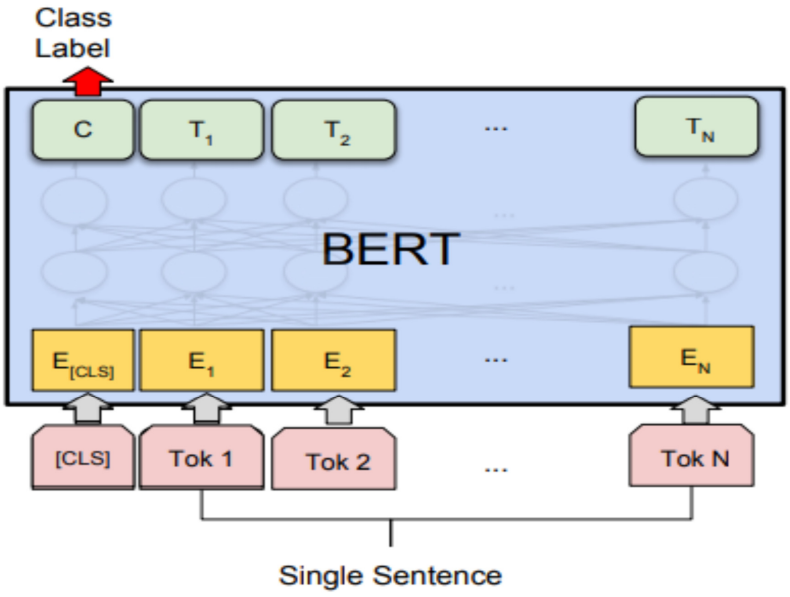

Fig. 2. BERT Model with Single Sentence

10) *Surprisal* (or negative log-conditional probability) tells us how strongly a certain word is expected under the language model's probability distribution (Levy, 2008).

In Lee (2021) and this paper, the final output part was revised so that the model could produce two values: (i) a class label (TRUE or FALSE) and (ii) the probability that the given sentence would be TRUE (acceptable). After the probability of TRUE was obtained for each input sentence, the values were normalized with both minimal and maximal values in the given dataset. Because clearly acceptable sentences and clearly unacceptable sentences were included in the filler sentences, all the values were located between 0 and 1.[11] Then, these values were converted into the acceptability scores, which ranged from 0 to 100.

3.5 Evaluation

After the acceptability scores were measured for all the sentences in the dataset, the validity of the acceptability scores was evaluated with the filler sentences (a total of 200 sentences), as in Lee (2021). The evaluation proceeded in two separate steps.

In the first step, the performance of the $BERT_{LARGE}$ model was measured with the class labels (TRUE or FALSE). 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 the correct answers in the CoLA dataset. 97.5% of accuracy was obtained in this step. It implies that our $BERT_{LARGE}$ model predicted the acceptability scores of English sentences with 97.5% of accuracy.

In the second step, the predictions of acceptability scores were grouped into two types (TRUE or FALSE). If the predicted score was equal to or greater than 50, the sentence was assigned the label TRUE. Otherwise, the sentence had the label FALSE. Then, the converted labels were compared with the predicted labels of the $BERT_{LARGE}$ model. 98.0% of accuracy was obtained in this second step. It implies that the $BERT_{LARGE}$ model predicted the class labels of sentences (acceptable or unacceptable) with 98.0% of accuracy if the acceptability scores were converted into two labels (grammatical or ungrammatical). The 2.0% of errors might occur during the normalization process. From these two separate steps of evaluation, it could be concluded that the predicted acceptability scores were 95.55% of accuracy for the target (island) sentences ($0.975 \times 0.980 = 0.9555$).[12]

4. Analysis Results

4.1 Descriptive Statistics

After all the acceptability scores were collected for both experimental and deep learning approaches, the scores were converted into *z*-scores. Then, the acceptability scores for two different complex NP island constraints were analyzed with the plots. The followings were the analysis results.

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

12) The accuracy (95.55%) was slightly lower than the value in Lee (2021). In Lee (2021), 98.25% of accuracy was obtained in the first step, and 97.75% of accuracy was obtained in the second step. As a result, about 96% of accuracy was expected for the target (island) sentences ($0.9825 \times 0.9775 = 0.9604 > 0.9555$). A slight reduction of the accuracy would be due to the size of the filler sentences. A total of 400 filler sentences were used in Lee (2021), but only half of them (200 filler sentences) were used in this paper. Note, however, that the accuracy of 95.55% was much higher than the accuracy of the LSTM model (fn. 2). From the comparisons, it can be said that the BERT model had the better performance.

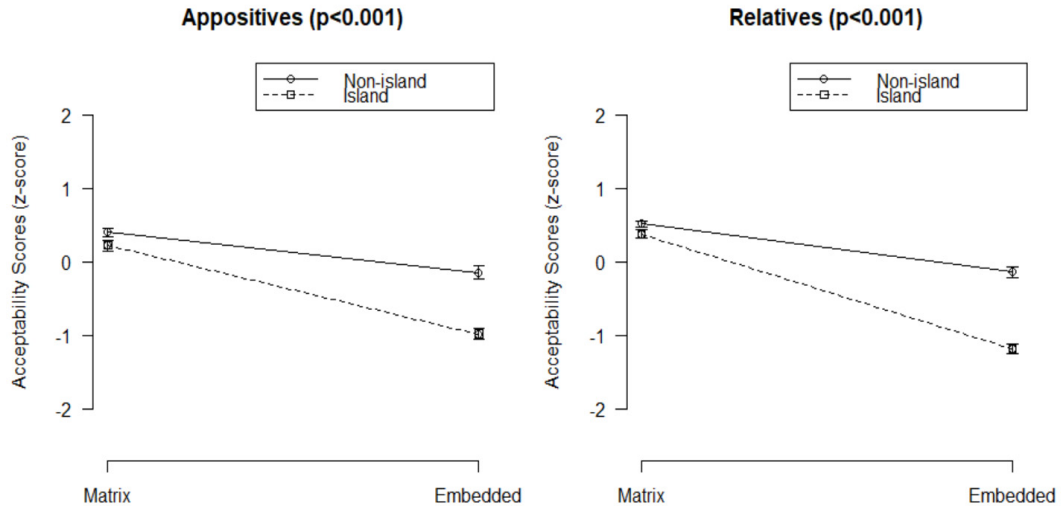

Fig. 3. Analysis Results in the Experimental Approach

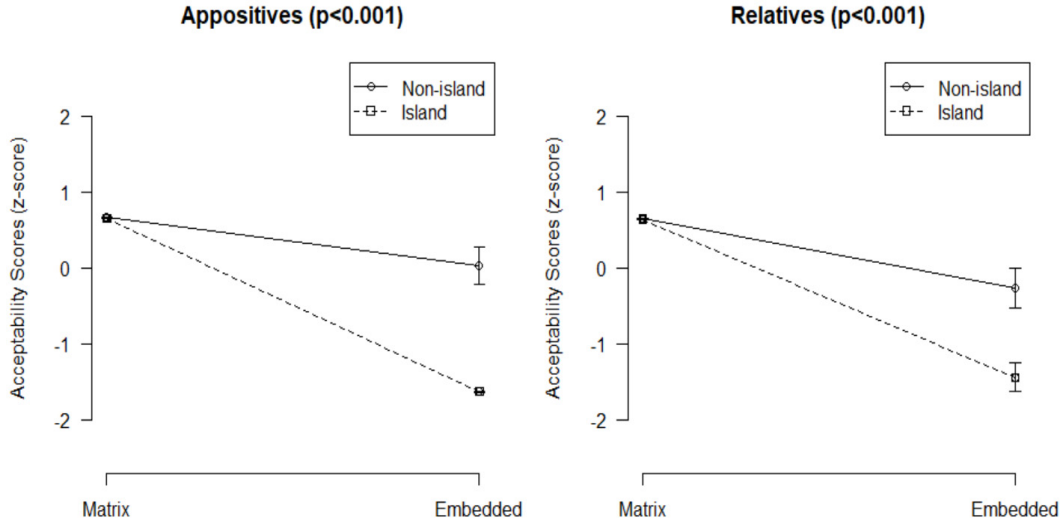

Fig. 4. Analysis Results in the Deep Learning Approach

If you compared two plots with those in Figure 3 and Figure 4, you could find that the overall tendency of the two plots was identical. It implied that the deep learning model correctly identified the island constraints in English. Two important differences were (i) that the confidence intervals (CIs) in Figure 4 were smaller or bigger than those in Figure 3 whereas the CIs in Figure 3 were similar, and (ii) that the difference between [Non-island, Embedded] and [Island, Embedded] was bigger in Figure 4.

4.2 Regression Analysis: Experimental Data

In order to examine whether two linguistic factors (*Island* and *Location*) influenced the acceptability scores of island

sentences, this study also conducted regression analyses for the experimental data. When the normality tests (Lee, 2016) were conducted to the *z*-scores, it was found that most of the data sets did not follow the normal distribution. Accordingly, a Generalized Linear Model (GLM) had to be used with a Gaussian distribution (a non-parametric test). The following two tables were the analysis results.

Table 1. GLM Analysis Results (Appositives)

	Estimate	*sd*	*t*	*p*
(Intercept)	-0.127	0.019	-6.512	<.001 ***
Location	-0.438	0.019	-23.073	<.001 ***
Island	-0.254	0.019	-13.370	<.001 ***
Location:Island	-0.163	0.019	-8.573	<.001 ***

Table 2. GLM Analysis Results (Relatives)

	Estimate	*sd*	*t*	*p*
(Intercept)	-0.106	0.016	-6.756	<.001 ***
Location	-0.555	0.016	-35.262	<.001 ***
Island	-0.298	0.016	-18.915	<.001 ***
Location:Island	-0.225	0.016	-14.313	<.001 ***

Lee and Park (2018) mentioned that both linguistic factors (*Island* and *Location*) independently affected the acceptability scores of island sentences (p<.001), in both Appositives and Relatives. In addition, the interaction between two factors (*Island*:*Location*) also influenced the acceptability scores of island sentences (p<.001).

4.3 Regression Analysis: Deep Learning Data

The same regression analyses were also applied to the deep learning data, in order to examine if two factors (*Island* and *Location*) influenced the acceptability scores of island sentences. Because the number of target sentences was small in the deep learning data set and most of the data sets did not follow the normal distribution, a GLM had to be used with a Gaussian distribution (a non-parametric test) in the analysis of data. The following two tables were the analysis results.

Table 3. GLM Analysis Results (Appositives)

	Estimate	*sd*	*t*	*p*
(Intercept)	-0.068	0.031	-2.203	.029 *
Location	-0.729	0.031	-23.769	<.001 ***
Island	-0.418	0.031	-13.636	<.001 ***
Location:Island	-0.415	0.031	-13.537	<.001 ***

Table 4. GLM Analysis Results (Relatives)

	Estimate	*sd*	*t*	*p*
(Intercept)	-0.103	0.040	-2.557	.011 *
Location	-0.752	0.040	-18.589	<.001 ***
Island	-0.295	0.040	-7.285	<.001 ***
Location:Island	-0.289	0.040	-7.144	<.001 ***

As you could observe from these tables, both factors (*Island* and *Location*) independently affected the acceptability scores of island sentences (p<.001) as in the experimental dataset, in both Appositives and Relatives. In addition, the interaction between two factors (*Island*:*Location*) also affected the acceptability scores of island sentences (p<.001).

4.4 DD Scores

Sprouse et al. (2012) used the differences-in-differences (DD) scores (Maxwell and Delaney, 2003) to measure the strength of island effects. The DD scores were calculated as follows.

(8) Calculation of DD Scores
   a. D1 = NON-ISLAND|EMBEDDED - ISLAND|EMBEDDED
   b. D2 = NON-ISLAND|MATRIX - ISLAND|MATRIX
   c. DD = D1 -D2

They mentioned that the plus values in the DD scores were *super-additive* effects. As you could observe in Figure 1, The differences between the scores in the embedded sentences were much greater than those in the matrix clauses. These effects clearly indicated the existence of island effects in English sentences.

This paper followed the method in Sprouse et al. (2012) and calculated the DD scores for both Appositives and Relatives. The following two tables enumerate the DD scores in both experimental and deep learning approaches.

Table 5. DD Scores (Experiment)

Island Type	Type 1	Type 2	Type 3	Type 4	D1	D2	DD
Appositives	0.405	-0.145	0.223	-0.978	0.833	0.182	0.651
Relatives	0.521	-0.138	0.376	-1.184	1.046	0.145	0.901

Table 6. DD Scores (Deep Learning)

Island Type	Type 1	Type 2	Type 3	Type 4	D1	D2	DD
Appositives	0.664	0.037	0.658	-1.630	1.666	0.006	1.660
Relatives	0.655	-0.272	0.643	-1.440	1.168	0.011	1.157

As you could observe in these tables, the DD score of Relatives was bigger than that of Appositives in the experimental approach. In the deep learning approach, the DD score of Relatives was smaller than that of Appositives.

It implied that Relatives were observed to have a stronger island effect than Appositives in the experimental data, and the opposite tendency was observed in the deep learning data.

5. Discussion

5.1 Differences between Experiments and Deep Learning

In this paper, the $BERT_{LARGE}$ model was combined with the CoLA dataset, and the acceptability scores for two different complex NP island constraints (Appositives and Relatives) were calculated with the pretrained deep learning model. After all the acceptability scores were obtained from two different kinds of approaches (an experiment and a deep learning model), statistical analyses were employed to compare the acceptability scores of the experiment and those of the $BERT_{LARGE}$ model.

As you could see in the comparison of Figure 3 and Figure 4, the overall tendencies were very similar in both sorts of approaches. In the experimental results and the $BERT_{LARGE}$ model, the island effects (the *super-additive* effects) were observed both in Appositives and Relatives. This result implied that the deep learning model (the $BERT_{LARGE}$ model) could correctly capture humans' syntactic acceptability, which supported Goldberg's claim (2019). That is, it implied that it was possible to employ deep learning models in the investigation of syntactic phenomena.

There were, however, some differences between the results of the deep learning models and those of the experiment. Note that the CIs in the deep learning data were bigger or smaller than those in experimental data but that the CIs in experimental data were very similar to one another. These discrepancies came from the number of the data. In the experimental data, each type of construction (each point in Figure 3) contained 100 data. In the deep learning data, each type of construction (each point in Figure 4) included only 5 values. Since the CIs decrease if the number of data points increases, it is natural that the CIs in the experimental data is smaller than those in the deep learning data. Then, what implications do the CIs have in the deep learning data? In the deep learning data, note that the 5 sets of sentences were constructed by only the lexicalization (i.e., by changing the words in the sentences but remaining the constructions unchanged). Then, the CIs in the deep learning data indicated the effects of lexicalzation (a random factor in a sense of Barr et al. (2013)). Note that three CIs were much bigger in the deep learning data: [Non-island, Embedded] in Appositives, [Non-island, Embedded] in Relatives, and [Island, Embedded] in Relatives. The others had very small CIs. This implied that the acceptability scores in the above 3 cases could be heavily influenced by the choice of lexical items. This kind of lexical influence could also be observed in the experimental data. However, since two random factors (participants and lexicalization) were combined in the experimental data, it was NOT easy to isolate only the effects of the lexicalization in the syntactic experiments.

Also note that the distances between [Non-island, Embedded] and [Island, Embedded] were wider in the deep $BERT_{LARGE}$ model than in the experimental data. As pointed out in Lee (2021), these bigger differences in the deep learning data could be originated from the Rectified Linear Unit (ReLU) function which was applied at the end of the process (when the $BERT_{LARGE}$ model made outputs for the [CLS] labels).[13] Since the ReLU function might maximize

13) For details on ReLU, see Agarap (2018).

the differences among the scores around the possibility of 0.5, it made the differences in the embedded clauses bigger in the deep learning model.

There were no typical things to be mentioned as for the GLM tables in Section 4.2 and those in Section 4.3. As you have already observed, "both factors (*Island* and *Location*) and their interaction (*Island*:*Location*) were statistically significant (p<.001). It implied that these two linguistic factors and their interaction influenced the acceptability scores of island sentences." These results demonstrated that the $BERT_{LARGE}$ model correctly identified the complex NP island constraints.

As for the DD scores, the DD score of Relatives was bigger than that of Appositives in the experimental approach. In the deep learning approach, on the other hand, the DD score of Relatives was smaller than that of Appositives. Since the DD scores indicated the strength of the island constraints (Sprouse et al., 2012; Sprouse and Hornstein, 2013), it implied that Relatives had stronger effects than Appositives in the experimental approach, and the opposite tendency was observed in the deep learning approach. This result had an implication that more studies were necessary to determine (i) whether Relatives were much stronger islands than Appositives or the opposite was true, and (ii) whether the complex NP islands were strong islands or weak islands à la Szabolcsi (2007).

5.2 Implications on Experimental Designs

Lee (2021) mentioned some implications of using deep learning technology in the experimental design of syntactic phenomena in natural languages. The same implications could be applied here. First, we do not need to worry about the fatigue of human experiments in the syntactic experiments. Second, if we conduct the experiments with the help of the deep learning data, we can try various lexical items and examine the influences of the lexical items. Third, an experiment with a deep learning model will be useful when we want to estimate the scores of sentences which were not included in the actual human experiments.

One more thing that we have to consider is that we isolate one random factor (lexicalization) in the deep learning data, whereas the experimental data with human participants usually have at least two random factors (variations speakers/individuals and variations in lexical items). This implies that we may have some changes in the random effects structure if we tried to analyze the data with a mixed-effects model (Barr et al., 2013).

Lee (2021) and this paper, however, do NOT claim that the deep learning model(s) such as the $BERT_{LARGE}$ model in this paper can substitute the experimental designs. What Lee (2021) and this paper want to say is that the deep learning models can be used in the study of syntax for compensating some shortcomings of experimental designs. In experimental syntax, the shortcomings come from the limitations of experimental designs including the number of participants or individual variations.

6. Conclusion

In this paper, a deep learning technique was employed and acceptability scores were calculated for two complex NP island constraints in English (Appositives and Relatives). This paper adopted the $BERT_{LARGE}$ model which was pretrained with the CoLA dataset. After the dataset was constructed with the sentences (some from Lee and Park (2018)

and others from the CoLA dataset), the acceptability scores were calculated for all the sentences both in the experimental approach and the deep learning approach (the $BERT_{LARGE}$ model). After the acceptability scores were measured, the scores were normalized into *z*-scores and statistical analyses were conducted.

The analysis results illustrated that the $BERT_{LARGE}$ model clearly identified two complex NP island constraints in English (Appositives and Relatives). Through the analysis, the followings were observed: (i) both the experimental approach and the $BERT_{LARGE}$ model correctly identified two complex NP island constraints in English, (ii) two factors (*Island* and *Location*) and their interaction (*Island*:*Location*) affected the acceptability scores of island sentences, and (iii) two approaches made different predictions on the DD scores of the two complex NP island constraints.

This paper demonstrated how deep learning technology could help the experimental design in syntax and how deep learning could be used in the experimental design. I hope that the combination of deep learning and experimental design can reveal a new way in the experimental syntax.

References

Agarap, F. 2018. Deep Learning Using Rectified Linear Units (ReLU). arXiv Preprint arXiv:1803.08375.

Alexopoulou, T. and F. Keller. 2007. Locality, Cyclicity, and Resumption: At the Interface between the Grammar and the Human Sentence Processor. *Language* 83, 110-160.

Annamoradnejad, I. and G. Zoghi. 2020. ColBERT: Using BERT Sentence Embedding for Humor Detection. arXiv preprint arXiv:2004.12765.

Bard, E., D. Robertson, and A. Sorace, 1996. Magnitude Estimation of Linguistic Acceptability. *Language* 72, 32-68.

Barr, D., R. Levy, C. Scheepers, and H. Tily. 2013. Random Effects Structure for Confirmatory Hypothesis Testing. *Journal of Memory and Language* 68, 255-278.

Carnie, A. 2021. *Syntax*: *A Generative Introduction*. Oxford: Wiley Blackwell.

Charniak, E., D. Blaheta, N. Ge, K. Hall, J. Hale, and M. Johnson. 2000. BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43. Philadelphia, PA: Linguistic Data Consortium.

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

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

Chomsky, N. 2000. Minimalist Inquiries: The Framework. In R. Martin, D. Michaels, and J. Uriagereka (eds.), *Step by Step*: *Essays on Minimalist Syntax in Honor of Howard Lasnik*. Cambridge, MA: MIT Press, 89-157.

Cowart, W. 1997. *Experimental Syntax*: *Applying Objective Methods to Sentence Judgments*. Thousands Oaks, CA: Sage Publications.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 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., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. arXiv preprint arXiv:1803.11138.

Hagstrom, P. 1998. *Decomposing Questions*. Cambridge, MA: MIT dissertation.

Hofmeister, P. and I. Sag. 2010. Cognitive Constraints on Syntactic Islands. *Language* 86, 366-415.

Hu, J., S. Chen, and R. Levy. 2020a. A Closer Look at the Performance of Neural Language Models on Reflexive Anaphor Licensing. *Proceedings of the Society for Computation in Linguistics*, 323-333.

Hu, J., J. Gauthier, P. Qian, E. Wilcox, and R. Levy. 2020b. A Systematic Assessment of Syntactic Generalization in Neural Language Models. arXiv preprint arXiv:2005.03692.

Keller, F. 2000. *Gradient in Grammar*: *Experimental and Computational Aspects of Degrees of Grammaticality*. 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, and B. Schmeiser (eds.), *Proceedings of the West Coast Conference on Formal Linguistics 23*. Somerville, MA: Cascadilla Press, 475-499.

Kluender, R. and M. Kutas. 1993. Subjacency as a Processing Phenomenon. *Language and Cognitive Processes* 8, 573-633.

Lasnik, H. and M. Saito. 1984. On the Nature of Proper Government. *Linguistic Inquiry* 15, 235-289.

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, 23-47.

Lee, Y. and Y. Park. 2018. English Island Constraints by Natives and Korean Non-natives. *The Journal of Studies in Language* 34, 439-455.

Levy, R. 2008. Expectation-based Syntactic Comprehension. *Cognition* 106, 1126-1177.

Marvin, R. and T. Linzen. 2018. Targeted Syntactic Evaluation of Language Models. arXiv preprint arXiv:1808.09031.

Maxwell, S. and H. Delaney. 2003. *Designing Experiments and Analyzing Data*: *A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Park, K., M. Park, and S. Song. 2021. Deep Learning Can Contrast the Minimal Pairs of Syntactic Data. *Linguistic Research* 38, 395-424.

Park, Y. and Y. Lee. 2018. English Island Sentences by Korean EFL Learners. *English Language and Linguistics* 24, 153-172.

R Core Team. 2021. 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*. 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. 2008. Magnitude Estimation and the Non-Linearity of Acceptability Judgments. In N. Abner and J. Bishop (eds.) *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, 397-403.

Sprouse, J. and N. Hornstein. 2013. *Experimental Syntax and Island Effects*. Cmabridge, MA: Cambridge University Press.

Sprouse, J., M. Wagers, and C. Phillips. 2012. A Test of the Relation between Working Memory Capacity and Syntactic Island Effects. *Language* 88, 82-123.

Szabolcsi, A. 2007. Strong vs. Weak Islands. In M. Everaert and H. van Riemsdijk (eds.), *The Blackwell Companion to Syntax*. Oxford: Blackwell, 479-531.

Szabolcsi, A. and F. Zwarts. 1993. Weak Islands and an Algebraic Semantics of Scope Taking. *Natural Language Semantics* 1, 235-284.

Truswell, R. 2007. Extraction from Adjuncts and the Structure of Events. *Lingua* 117, 1355-1377.

Tsai, W. 1994. On Nominal Islands and LF Extraction in Chinese. *Natural Language and Linguistic Theory* 12, 121-175.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv: 1804.07461.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General- purpose Language Understanding Systems. arXiv preprint arXiv:1905.00537.

Warstadt, A., Y. Cao, I. Grosu, W. Peng, H. Blix, Y. Nie, A. Alsop, S. Bordia, H. Liu, A. Parrish, S. Wang, J. Phang, A. Mohananey, P. Htut, P. Jeretič, and S. Bowman. 2019a. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. arXiv preprint arXiv:1909.02597.

Warstadt, A., A. Singh, and S. Bowma. 2019b. Neural Network Acceptability Judgments. arXiv preprint arXiv:1805.12471.

Wilcox, E., R. Levy, and R. Futrell. 2019a. What Syntactic Structures Block Dependencies in RNN Language Models? arXiv preprint arXiv:1905.10431.

Wilcox, E., R. Levy, and R. Futrell. 2019b. Hierarchical Representation in Neural Language Models: Suppression and Recovery of Expectations. arXiv preprint arXiv:1906.04068.

Wilcox, E., R. Levy, T. Morita, and R. Futrell. 2018. What Do RNN Language Models Learn about Filler-Gap Dependencies? arXiv preprint arXiv:1809.00042.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *Proceedings of the IEEE International Conference on Computer Vision*, 19-27.

Lee, Yong-hun, Instructor
99 Daehak-ro, Yuseong-gu, Daejeon 34134, Korea
Department of Linguistics Chungnam National University
E-mail: yleeuiuc@hanmail.net