Korean Journal of English Language and Linguistics, Vol 23, June 2023, pp. 482-497
DOI: 10.15738/kjell.23..202306.482

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

On Pronoun Prediction in the L2 Neural Language Model*

Sunjoo Choi · Myung-Kwan Park (Dongguk University)

Received: May 25, 2023
Revised: June 5,2023
Accepted: June 23, 2023

Sunjoo Choi (first author)
Post-Doctor, Division of English Language and Literature, Dongguk University
Email: sunjoo@dongguk.edu

Myung-Kwan Park (corresponding author)
Professor, Division of English Language and Literature, Dongguk University
Tel: 82-2-2260-8708
Email:parkmk@dongguk.edu

* This work was supported by the Dongguk University Research Fund of 2022 (S-2022-G0001-00134).

ABSTRACT

Choi, Sunjoo and Myung-Kwan Park. 2023. On pronoun prediction in the L2 neural language model. *Korean Journal of English Language and Linguistics* 23, 482-497.

In recent years, artificial neural(-network) language models (LMs) have achieved remarkable success in tasks involving sentence processing. However, despite leveraging the advantages of pre-trained neural LMs, our understanding of the specific syntactic knowledge acquired by these models during processing remains limited. This study aims to investigate whether L2 neural LMs trained on L2 English leaners' textbooks can acquire syntactic knowledge similar to that of humans. Specifically, we examine the L2 LM's ability to predict pronouns within the framework of previous experiments conducted with L1 humans and L1 LMs. Our focus is on pronominal coreference, a well-studied linguistic phenomenon in psycholinguistics that has been extensively investigated. This research expands on existing studies by exploring whether the L2 LM can learn Binding Condition B, a fundamental aspect of pronominal agreement. We replicate several previous experiments and examine the L2 LM's capacity to exhibit human-like behavior in pronominal agreement effects. Consistent with the findings of Davis (2022), we provide further evidence that, like L1 LMs, the L2 LM fails to fully capture the range of behaviors associated with Binding Condition B, in comparison to L1 humans. Overall, neural LMs face challenges in recognizing the complete spectrum of Binding Condition B and are limited to capturing aspects of it only in specific contexts.

KEYWORDS

neural language model, binding condition, pronoun, surprisal, coreference

1. Introduction

Artificial neural(-network) language models (LMs) have recently achieved notable accomplishments in investigating the connection between human linguistic representations and neural LM representations. Extensive research has been conducted to compare neural LMs with human linguistic behavior. A substantial body of literature has emerged, drawing inspiration from investigations in theoretical linguistics, psycholinguistics, and neurolinguistics, addressing these comparative analyses. Notably, many prior studies have concentrated on assessing the capacity of neural LMs to replicate human-like subject-verb agreement patterns (Gulordava et al., 2018, Linzen and Leonard 2018, Wilcox et al., 2018, Warstadt et al., 2020). Several other syntactic phenomena have been explored, and then neural LMs have been claimed to perform human-like behavior in processing syntactic islands (Wilcox et al., 2019), garden path constructions (van Schijndel and Linzen 2018), negative polarity items (Marvin and Linzen 2018, Jumelet and Hupkes 2018), and filler-gap dependencies (Bhattacharya and van Schijndel 2020).

Given the predominant focus of previous studies on sentence processing performances by L1 LMs, the present study shifts its attention towards L2 LMs trained on L2 English learners' textbooks (henceforth, for brevity, L2 LMs). Our objective is to investigate whether these L2 LMs, trained on the data that Korean L2 learners utilize to learn English, can exhibit human-like behavior in sentence processing. Through this inquiry, we aim to present new evidence regarding the L2 LM's potential to acquire linguistic systems resembling those of native speakers. To accomplish this, we specifically examine a well-established linguistic phenomenon in psycholinguistics: pronominal coreference associated with binding. In particular, we explore the interplay between Binding Condition B and coreference processing. By doing so, we can evaluate the extent to which the L2 LM's pronoun prediction aligns with Binding Condition B, enabling us to assess the LM's syntactic knowledge. The provided example in (1) serves as a pertinent illustration of our investigation.

(1) Mary explained to John that Sue had deceived her.

In (1), despite *her* agreeing in gender with both *Mary* and *Sue*, *her* unambiguously refers to *Mary*. This is a well-known fact that Binding Condition B (Chomsky 1981) blocks *her* from referring to *Sue*. In a nutshell, Binding Condition B demonstrates that a pronoun must be free within its local domain. In this sense, if *her* refers to Sue, then *her* would be bound in its local domain (that is, resulting in a violation of Binding Condition B). Comparing cues for pronominal antecedents has been a central issue of experimental work. These cues are often considered as linguistic constraints. In general, these cues are divided into two classes: those that make reference to the agreement features on the pronoun (e.g., gender, number, person) and those that make reference to structural constraints (e.g., Binding Conditions). Several previous works have found that structural constraints immediately constraint the set of possible antecedents (e.g., Clifton et al., 1997, Chow et al., 2014, Kush and Dillon 2021). Such works would suggest that the initial set of possible antecedents is *Mary* and *John*, with the gender of *her* excluding *Sue* later. However, other works suggested that the initial set may also contain *Sue*. It means that illicit antecedents also have measurable effects (e.g., Badecker and Straub 2002, Kennison 2003).

In a recent study, Davis (2022) investigated whether neural LMs possess syntactic constraints concerning coreference. By extending previous research on Binding Condition B, Davis examined multiple neural LMs. While it remains challenging to assess whether neural LMs can accurately interpret pronouns as coindexed with specific antecedents, it is possible to compare the behavioral distinctions among these models using subtly different stimuli. Davis replicated human experiments using neural LMs of English, specifically L1 neural LMs, in order to facilitate the comparison. This prior work offered compelling evidence that the models exhibited a stronger attraction towards antecedents for pronouns compared to humans, while still demonstrating behavior in line with Binding Condition B. However, as the experiments progressed, the L1 LMs struggled to capture the intricate interactions between Condition B and other linguistic processes. In the next section, we will delve into a comprehensive review of Davis's experimental study.

2. Previous study: Davis (2022)

Recently, Davis (2022) examined how neural LMs predict pronouns by building upon previous experimental research conducted with native English speakers. The primary objective of his work was to investigate whether linguistic constraints are assessed simultaneously or in a particular order. Specifically, he explored whether neural LMs have the ability to attend to gender agreement between pronouns and potential antecedents before considering Binding Condition B, or if the structural constraint imposed by Binding Condition B takes precedence. To illustrate this inquiry, let us consider the following four examples.

(2) a. Mary thought Olivia hated her.
b. Mary thought John hated her.
c. Austin thought Olivia hated her.
d. Austin thought John hated her.

Given these examples, if gender agreement is initially applied to restrict the candidate set, we anticipate that (2a), (2b), and (2c) would exhibit a similar pattern as they all have at least one potential antecedent, while example (2d) lacks any candidate sets. However, if Binding Condition B and agreement are interconnected, we expect (2a) and (2b) to pattern together because they both have one possible antecedent, namely ‘Mary’. Similarly, (2c) and (2d) would pattern together as they yield no possible antecedents. In simpler terms, the presence of the embedded subject (*Olivia* or *John*) influences behaviors towards it, depending on whether Binding Condition B operates earlier or later during pronoun processing. Previous studies have revealed that structural constraints promptly limit the pool of potential antecedents (Clifton et al., 1997, Chow et al., 2014, Kush and Dillon 2021). Based on these findings, it can be observed that (2a) and (2b) demonstrate a similar pattern, while (2c) and (2d) exhibit a comparable pattern as well. However, other studies have suggested that grammatically illicit antecedents can still have measurable effects (Badecker and Straub 2002, Kennison 2003), indicating that the initial set of candidates may include ungrammatical options. It is important to note that task-specific environments can influence different candidate sets or capture different processing stages. Considering the previous findings, it is evident that Binding Condition B has an immediate impact on processing, with potential subsequent repair mechanisms.

To extend the existing studies of neural LMs to the investigation into Binding Condition B, Davis employed the 25 LSTM models trained on Wikitext-103 (Merity et al., 2016) and BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), Transformer XL (Dai et al., 2019), and GPT-2 XL (Radford et al. 2019). To assess the probability assigned to a pronoun, he adopted surprisal (Hale 2001, Levy 2008)[1], which is defined as follows: -log P (word | context). Furthermore, using surprisal values, a gender mismatch effect is calculated[2]. The stimuli of the first experiment are drawn from Chow et al. (2014). There are 60 sets of stimuli contrasting whether the main subject matches the embedded object pronoun in gender and whether the embedded subject matches the embedded object pronoun in gender (that is, 2 x 2 factorial design). An example set is given below.

(3) a. Martin dreamed that the wizard would poison him surreptitiously on the night of the full moon. (Match,

[1] Surprisal values have been claimed to linearly correlate with human reading times (Hale 2001, Levy 2008, Smith and Levy 2013).

[2] In human experiments, gender mismatch effects mean the increased cost in processing a pronoun with an unexpected gender. In the experiment with neural LMs, Davis calculated gender mismatch effects for both pronoun prediction and antecedent prediction.

Match)

b. Martin dreamed that the witch would poison him surreptitiously on the night of the full moon. (Match, Mismatch)

c. Brenda dreamed that the wizard would poison him surreptitiously on the night of the full moon. (Mismatch, Match)

b. Brenda dreamed that the witch would poison him surreptitiously on the night of the full moon. (Match, Mismatch) (Davis 2022: 124)

In addition, Davis added an additional condition for possessive pronoun *his*. For instance, examples like (3) generate additional stimuli of the following form: *Martin dreamed that the wizard would poison his lover surreptitiously on the night of the full moon*. As noted by Chow et al., for humans, (3a) and (3b) pattern together, as do (3c) and (3d). When examining the reading times associated with the pronoun *him*, it was observed that the pattern aligned with the gender of the matrix subject (*Martin* and *Brenda*), while no significant reading time differences were observed based on the gender of the embedded subject. Chow et al. (2014) interpreted this finding as evidence indicating that Binding Condition B promptly constrains the potential antecedent options for the embedded object *him*. Building upon these results, Davis identified and reported evidence supporting two distinct categories of neural LMs' behavior, as depicted in Figure 1.

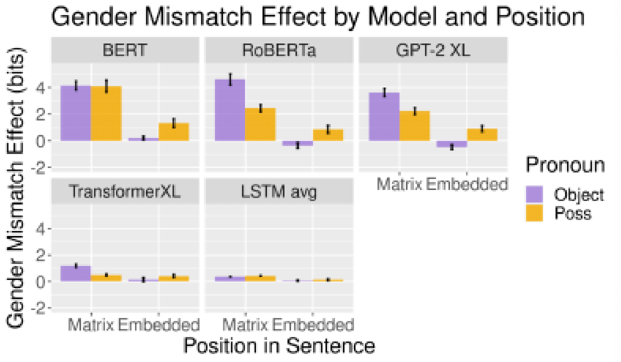

Figure 1. Gender Effect for Pronoun and Possessive Pronoun (Davis 2022: 125)

Davis's findings revealed two distinct patterns of behavior exhibited by the LMs. One group of LMs aligned with human behavior observed in Chow et al. (2014), where only the gender of the grammatically licit antecedent influenced the surprisal of the object pronoun *him*. The other group of LMs, including RoBERTa and GPT-2 XL, showed that both the gender of the matrix subject and the gender of the embedded subject influenced the surprisal of *him*. However, for models like BERT, Transformer XL, and the LSTMs, only the gender of the matrix subject affected the surprisal of *him*, indicating consideration of only grammatically licit positions for coreference. Regarding the pronoun *his*, all the models exhibited a consistent pattern, with the surprisal influenced by both the gender of the matrix subject and the gender of the embedded subject. Overall, Davis's findings demonstrated that when the matrix subject agreed in gender with it, the pronoun *him* was equally surprising, but when there was a lack of agreement, there was a cost associated with agreeing with the ungrammatical antecedent.

The findings regarding the pronoun *his* revealed that it was not subject to violations of Binding Condition B, but surprisal values were influenced by both the matrix and the embedded subject. In this case, agreement with the subject was preferred, indicating that LMs tend to favor agreement with the matrix subject. Interestingly, the divergent behavior between *his* and *him* suggests that neural LM behaviors for *him* are specific to that particular pronoun and not a general behavior for coreference.

For LMs where the surprisal of *him* was solely influenced by the matrix subject, the models disregarded ungrammatical antecedents in accordance with Binding Condition B. This pattern indicates some level of knowledge of Binding Condition B or, at the very least, a stronger preference for object pronouns. On the other hand, the pronoun *his* should be able to agree with either noun, but only the matrix subject demonstrated a noticeable influence on surprisal. The distinction observed between *his* and *him* in L1 LMs rules out a simple preference for *his* and suggests that this behavior is not solely governed by a heuristic tied to the pronoun *him*. This aligns with the similar human results reported by Chow et al. (2014) that followed a simple heuristic.

To address these inferential issues, Davis conducted a second experiment to differentiate general subject preferences from the behavior dictated by Binding Condition B. The first experiment involved two noun phrases (e.g., *NP1 said NP2 liked him*). However, taking inspiration from Nicol and Swinney (1989), Davis expanded the number of possible antecedents to three, including two grammatically plausible antecedents and one that violated Binding Condition B. The specific example is provided below:

(4) *The doctor* told *the lord* that *the uncle* would teach *him* how to drive this weekend.

According to Nicol (1988), the pronoun *him* specifically reactivated the antecedents *the doctor* and *the lord*. In other words, *him* reactivated both of the syntactically viable noun phrases, while blocking the activation of *the uncle* in accordance with Binding Condition B. The experiment consisted of 24 stimuli encompassing various gender combinations for the three noun phrases. The results of this experiment are depicted in Figure 2.

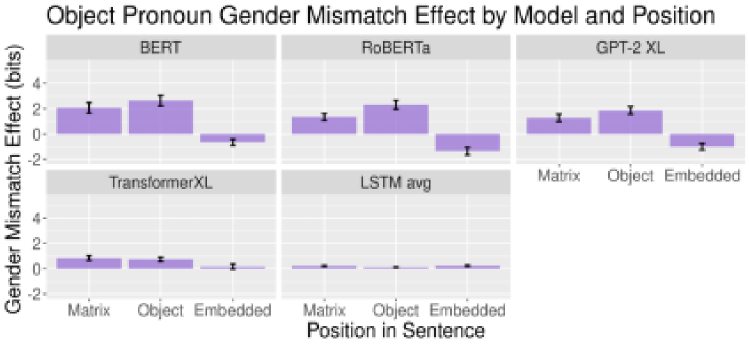

Figure 2. Gender Mismatch Effect for Object Pronoun (Davis 2022: 132)

In contrast to the first experiment, Davis observed more diverse behavior in the neural LMs when three noun phrases were involved. The LSTMs and TransformerXL did not exhibit interactions between antecedents; instead, the presence of masculine gender in any of the noun phrases reduced the surprisal of the pronoun *him*. On the other hand, the remaining models demonstrated a more intricate pattern with interactions among all three noun phrases.

Similar to the findings with two antecedents, Davis discovered that there was no penalty for agreement with ungrammatical antecedents when the grammatical antecedents agreed with the pronoun. Among GPT-2 XL, BERT, and RoBERTa, a behavior resembling a conditional constraint was observed. Contexts where the ungrammatical antecedent agreed with the pronoun were less preferred, indicating some correspondence with Binding Condition B.

Up until this point, the previous experiments have primarily focused on backward anaphora. The main objective of these two experiments was to investigate whether the gender of preceding nouns influences the gender of the pronoun. Building upon these findings, Davis further examined the interaction between Binding Condition B and the prediction of upcoming words. To confirm this, he explored whether neural LMs were constrained by cataphoric pronouns following Binding Condition B. Davis used stimuli adapted from van Gompel and Liversedge

(2003), which demonstrated that cataphoric pronouns influenced the prediction of subject nouns in human participants. Specifically, a masculine cataphoric pronoun led to a mismatch effect when the subject was feminine. The materials employed in the experiment were adopted from Experiment 1 in van Gompel and Liversedge (2003), as depicted in (5).

(5) a. When she was off work, the waitress pestered the waitress all the time.
b. When she was off work, the waiter pestered the waiter all the time.
c. When he was off work, the waitress pestered the waiter all the time.
d. When he was off work, the waiter pestered the waitress all the time.

In sentences such as (5a) and (5d), the cataphoric pronoun agrees in gender with the subject, while in (5b) and (5c) there is a mismatch. Van Gompel and Liversedge found that human readers tended to establish a connection between the pronoun and the subject. In such cases, reading times were prolonged when there was a gender mismatch. Based on these findings, Davis observed the patterns exhibited by the L1 LMs, which are illustrated in Figure 3.

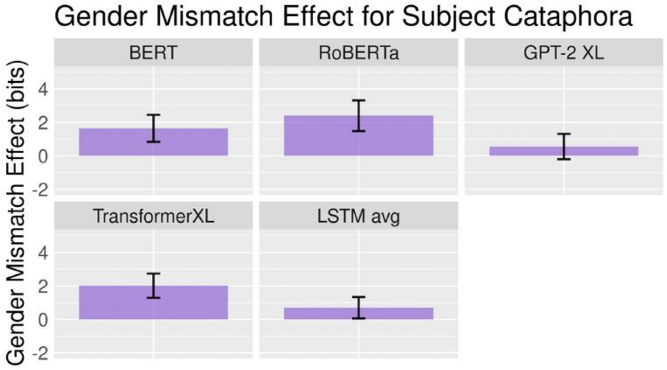

Figure 3. Gender Mismatch Effect for Subject Following a Cataphoric Subject Pronoun (Davis 2022: 138)

Davis presented new findings indicating that BERT, RoBERTa, and Transformer XL exhibited gender-dependent predictions for subjects following cataphoric pronouns. The LSTMs demonstrated constrained predictions specifically for masculine pronouns, while GPT-2 XL did not show any effects.

Thus far, Davis has obtained initial evidence regarding pronoun prediction based on preceding context. Interestingly, Kush and Dillon (2021) proposed that predictions could be influenced by whether Binding Condition B permits coreference between the cataphoric pronoun and the subject. Building on this suggestion, Davis replicated Kush and Dillon's experiments using L1 LMs. The stimuli were taken from Kush and Dillon (2021)[3], which examined whether Binding Condition B constrains cataphoric pronouns. Let us consider the following examples:

(6) a. Before offering *him* a fancy pastry, Michael politely asked Tyler whether he had any preference.
b. Before offering *his* son a fancy pastry, Michael politely asked Tyler whether he had any preference.

In sentence (6a), the pronoun *him* is unable to refer to *Michael* due to the effect of Binding Condition B. However,

[3] In Kush and Dillon's experiment, proper names were used. However, proper names were changed to *the man* or *the woman.*

in sentence (6b), coreference between *his* and *Michael* is possible and appears to be the more natural interpretation. Furthermore, a modified version of sentence (7) was used in a subsequent experiment.

(7) Before anyone offered *him* a fancy pastry, Michael politely asked Tyler whether he had any preference.

In sentence (7), the pronoun *him* is used, allowing for possible coreference with the subject *Michael*. Kush and Dillon conducted a study and found that there was a reading time slowdown when the cataphoric pronoun and the subject disagreed in gender, but this effect was observed only when Binding Condition B was not implicated. Building upon these findings, Davis examined to what extent neural LMs can capture this empirical generalization. A total of 24 stimulus sets were investigated in the study. Following the approach of Kush and Dillon (2021), both masculine and feminine pronouns were examined, and the results are depicted in Figure 4.

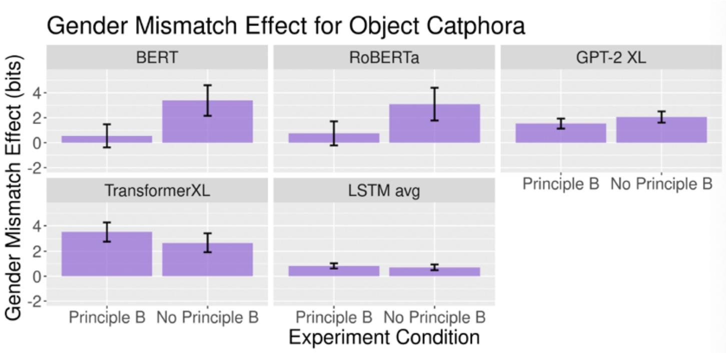

Figure 4. Gender Mismatch Effect for Subject Following a Cataphoric Object Pronoun (Davis 2022: 141)

These two experiments aimed to investigate the impact of Binding Condition B on the gender mismatch effect for cataphora. In human participants, a gender mismatch effect occurs when the subject can be grammatically co-indexed with the cataphoric pronoun. Interestingly, BERT and RoBERTa exhibited similar patterns to humans, while the other models did not. GPT-2 XL, TransformerXL, and the LSTMs only showed main effects of cataphora agreement, with higher surprisal when the subject and the cataphoric pronoun were mismatched.

In the second experiment conducted by Kush and Dillon (2021), they used the same pronoun forms (*him* and *her*) while manipulating the presence of Binding Condition B violations. In this case, RoBERTa and BERT did not demonstrate any effect of Binding Condition B. On the other hand, GPT-2 XL, TransformerXL, and the LSTMs displayed the same generalization as observed in the first experiment. They exhibited a mismatch effect in both conditions, irrespective of the presence of a Binding Condition B violation.

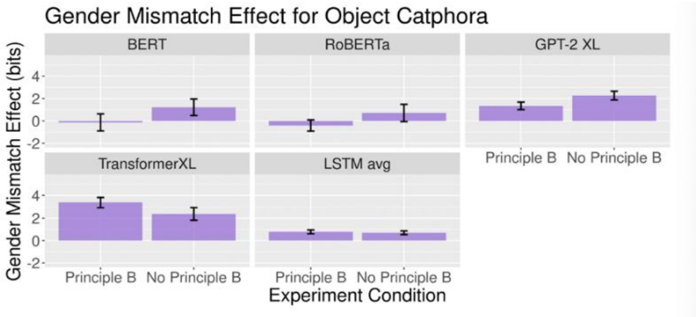

Figure 5. Gender Mismatch Effect for Subject Following a Cataphoric Object Pronoun (Davis 2022: 143)

Contrary to humans, the L1 LM's results demonstrated a gender mismatch effect when coreference between the subject and the pronoun was not blocked by Binding Condition B. L1 LMs learned to behave in accordance with Binding Condition B only in specific contexts. Across all these experiments, Davis did not find neural LM's regular behaviors in line with humans.

Considering L1 LM's behaviors, we turn to explore how the L2 LM influences the processing of pronouns. The present study straightforwardly extends the previous study to the L2 LM. Even though we cannot evaluate whether the L2 LM indeed processes the pronoun as coindexing with certain antecedents, we can directly compare the behavioral differences in both L1 LMs and humans using the same stimuli. In this sense, we begin by replicating four experiments in Davis's study with the LSTM LM for Korean L2 learners of English. This paper primarily focuses on the two research issues: (1) the extent to which the L2 LM predicts gender agreement between a pronoun and its possible antecedents, and (2) the extent to which it establishes coreference in accordance with Binding Condition B. In what follows, we introduce the architecture of the L2 LM and its pronoun and antecedent prediction in various linguistic environments.

3. Experimental Configurations

As the initial step of this study, we created the L2 Long-Short Term Memory (LSTM) language model. We adopted the Gulordava LSTM model, which was originally introduced by Gulordava et al. (2018) for learning English subject-verb number agreement. The Gulordava model was pretrained on 90 million tokens from English Wikipedia and consisted of two hidden layers, each comprising 650 units. Following the architecture of the Gulordava model, we constructed the L2 neural language model for our experiments.

To train the L2 LM, we collected training datasets from various sources, including the EBS-CSAT English Prep Books published between 2016 and 2018, as well as English textbooks used by Korean English L2 learners. These textbooks were based on the English 11 middle-school and L2 high-school textbooks published in Korea in 2001-2002, and the English 19 middle-school and L2 high-school textbooks published in Korea in 2009-2010 (Kim 2020, Choi et al., 2021, Choi and Park, 2022a,b,c,d). This data collection process resulted in a corpus of 7.9 million tokens extracted from English textbooks for Korean L2 learners of English.

Additionally, we implemented a data augmentation algorithm using the textbook corpus to expand the training tokens further, resulting in an additional dataset of 5.1 million tokens. This augmentation strategy allowed us to increase the size of the training data. Ultimately, we assembled a training dataset of 13 million tokens to train the L2 LSTM language model.

Before presenting the findings, it is important to acknowledge that previous studies have suggested that neural LMs are capable of capturing aspects of Binding Condition A and reflexive anaphora (Warstadt et al., 2020, Hu et al., 2020). Recent evaluations of LMs indicate that these binding conditions can be learned from the training data. In our study, we employed an L2 LM trained on a corpus of 13 million tokens, which represents the English learning materials commonly encountered by Korean English learners. We opted for the L2 language model based on its demonstrated efficacy in acquiring various syntactic structures, such as filler-gap dependencies (Kim 2020), linguistic anomalies (Choi et al., 2021), the dative alternation (Choi and Park 2022a), relative clauses (Choi and Park 2022b), self-paced reading time (Choi and Park 2022c), and transitive alternation (Choi and Park 2022d).

4. Results

4.1 Binding Condition B with two NPs

In order to examine the grammatical constraints on coreference, our investigation focused on determining which antecedent noun phrases (NPs) are available at the moment the pronoun is encountered. The primary objective was to analyze the behaviors exhibited by the L2 LM and their influence on the pronoun. To facilitate a comparison between the L2 LM and its L1 counterparts, we utilized the experimental items and codes from Davis (2022).[4] We conducted experiments using 60 sets of stimuli, exploring scenarios where the gender of the pronoun aligned with either the subject of the main clause or the subject of the embedded clause, as depicted in (8).

(8) a. The *man* thought that the *waiter* would praise *him* enthusiastically for the amazing success of the event.
b. The *man* thought that the *party planner* would praise *him* enthusiastically for the amazing success of the event.
c. The *woman* thought that the *waiter* would praise *him* enthusiastically for the amazing success of the event.
d. The *woman* thought that the *party planner* would praise *him* enthusiastically for the amazing success of the event.

As shown in (8), the subjects were limited to *the man* or *the woman* due to vocabulary constraints. Additionally, we acknowledged the potential ambiguity of the feminine pronoun *her*, as it could be interpreted as an object pronoun subject to Binding Condition B. Therefore, we focused solely on the masculine pronoun, as suggested by Davis. Furthermore, we introduced an additional condition that compared *him* and *his*, as illustrated below.

(9) a. The *man* thought that the *waiter* would praise *his* co-worker for the success of the event.
b. The *man* thought that the *party planner* would praise *his* co-worker for the success of the event.
c. The *woman* thought that the *waiter* would praise *his* co-worker for the success of the event.
d. The *woman* thought that the *party planner* would praise *his* co-worker for the success of the event.

Figure 6 displays the results of the L2 LM. The statistical analysis was performed using a linear-mixed effects model. In all the graphs presented in Section 4, a positive value indicates that the pronoun gender was predicted to agree with the antecedent, while a negative value indicates a predicted disagreement between the pronoun gender and the antecedent. The error bars represent the 95% confidence intervals.

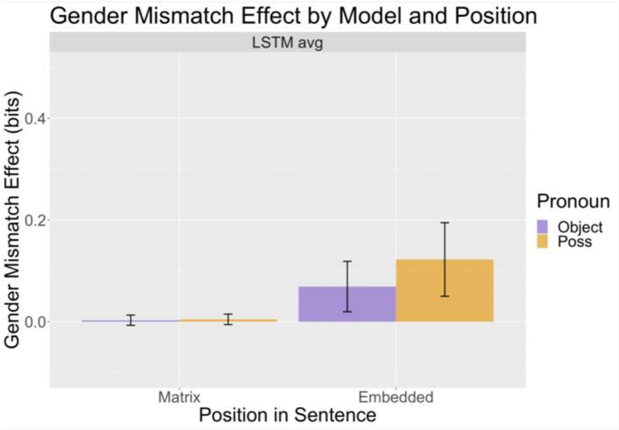

Figure 6. The L2 LM's Gender Mismatch Effect for Object Pronoun (*him*) and Possessive Pronoun (*his*)

[4] As pointed out in Section 2, Davis employed various neural LMs (BERT, RoBERTa, GPT-2 XL, TransformerXL, and LSTM) for the study. However, for our specific investigation, we focused on implementing the L2 LSTM LM, drawing inspiration from the Gulordava LSTM LM. In this context, we analyze and compare the patterns exhibited by the L2 LSTM LM with those of the L1 LSTM LM, without going into its comparison to other types of neural LM.

The L2 LSTM LM demonstrated different patterns compared to the L1 LSTM LMs. The gender of the embedded subject had a stronger influence on the surprisal of the pronoun *him* compared to the gender of the matrix subject, indicating a higher gender mismatch effect. This suggests that grammatically illicit positions were considered for coreference in the L2 LM. The L2 LM showed a lowered surprisal when the matrix subject was mismatched in gender feature with the pronoun. The L2 LM failed to capture what Binding Condition B dictates. The influence of the embedded subject on pronoun processing was more prominent, indicating measurable effects of grammatically illicit antecedents. To examine the status of Binding Condition B, we compared the gender mismatch effects for *him* and *his*. In the case of the pronoun *his*, similar to the pronoun condition, the L2 LM exhibited a stronger influence of the gender of the embedded subject on the surprisal of the pronoun. This effect was more pronounced in the embedded pronoun condition.

Overall, these combined results suggest that the L2 LM did not successfully learn the knowledge of Binding Condition B and instead demonstrated a stronger subject preference for object pronouns. The observed patterns in the gender of the embedded NPs indicate that Binding Condition B does not function as an initial filter for antecedents in the L2 LM. To gain deeper insights into the behavior of the LM, we aimed to distinguish between adherence to Binding Condition B and a simple subject preference. By doing so, we hoped to find further evidence regarding the knowledge of Binding Condition B in the L2 LM.

4.2 Binding Condition B with three NPs

Following the approach used in the experiments with the L1 neural LMs, we extended the number of possible antecedents from two to three in our experiments, as in (10).

(10) *The man* told *the wizard* that *the nephew* would protect *him* if it became necessary.

As mentioned earlier, the second experiment aimed to differentiate between adherence to Binding Condition B and a simple subject preference by introducing an additional NP. If the L2 LM solely considers antecedents following Binding Condition B, then the surprisal of the pronoun should be influenced by NP1 (*the man*) and NP2 (*the wizard*). On the other hand, if subject preference affects the behavior of the L2 LM, then only NP1 (*the man*) should impact the surprisal of the pronoun. In the context of this experiment, positive values indicate that the pronoun gender was predicted to agree with the antecedent, while negative values indicate disagreement.

The results obtained from the L2 LM are presented in Figure 7. Similar to the L1 LSTM LMs, the L2 LM demonstrated a general effect of masculine agreement. However, unlike the L1 LSTM LMs, the L2 LM exhibited an ungrammatical match effect in certain contexts. Overall, the coreference preferences of the L2 LM extend beyond a general preference for the first noun phrase (NP1). A conditional constraint was observed, where contexts in which the ungrammatical antecedent agreed with the pronoun were dispreferred. This pattern aligns with Binding Condition B. It is important to note that the influence of NP3 was not observed in humans, indicating that any preferences for NP3 are non-human-like behaviors. Instead, the observed patterns resemble those observed in humans and several L1 LMs (BERT, RoBERTa, and GPT-2 XL).

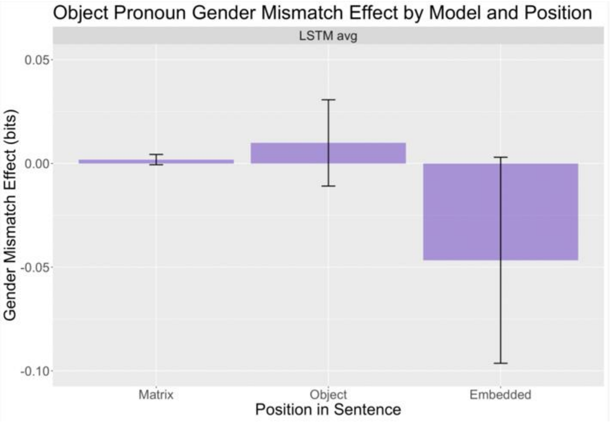

Figure 7. The L2 LM's Gender Mismatch Effect for Object Pronoun

At first glance, we have uncovered some evidence suggesting that the L2 LM exhibits behavior aligned with Binding Condition B. However, given the results from the first experiment, it remains uncertain whether the L2 LM possesses explicit knowledge of Binding Condition B. Therefore, it is necessary to investigate the interplay between Binding Condition B and the prediction of forthcoming material. Similar to Davis' approach, we will delve deeper into whether the L2 LM adheres to Binding Condition B by examining whether cataphoric pronouns constrain predictions in a manner consistent with Binding Condition B, as observed in humans. This exploration will allow us to ascertain whether the behavior of the L2 LM genuinely reflects the tracking of aspects related to Binding Condition B. This issue will be taken up in the next section.

4.3. Predictive Processing with Cataphora

As previously mentioned, the focus of the third experiment revolved around two key aspects of the L2 LM's linguistic behavior: (1) the anticipation of subject nouns based on the gender of the cataphoric pronoun, and (2) the linguistic knowledge pertaining to Binding Condition B. Within this subsection, we specifically examined the impact of a masculine cataphoric pronoun on the presence of a mismatch effect when the subject was feminine. Consistent with the previous experiments, we explored the interaction between Binding Condition B and the prediction of forthcoming material. To accomplish this, we utilized the stimuli from van Gompel and Liversedge's (2003) study, replicating their experiment and applying it to the L2 LM. Our aim was to investigate whether cataphoric pronouns impose restrictions on the expectation of upcoming nominal elements, addressing the aforementioned issues. Notably, in van Gompel and Liversedge's stimuli, pronouns preceded their antecedents, as depicted in (11) and reiterated from (4).

(11) a. When he was present, the actress embarrassed the actor all the time.
b. When he was present, the actor embarrassed the actress all the time.
c. When she was present, the actress embarrassed the actor all the time.
d. When she was present, the actor embarrassed the actress all the time.

As shown in Figure 8, no significant effects was in this experiment. However, this behavior was observed in GPT-2 XL, but not in L1 LSTMs.

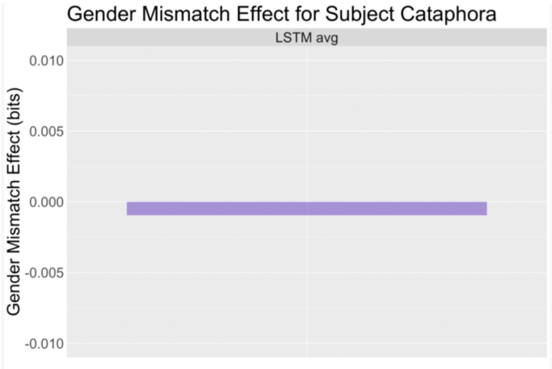

Figure 8. The L2 LM's Gender Mismatch Effect for Subject following a Cataphoric Subject Pronoun

Contrary to our expectations, the L2 LM did not demonstrate the ability to predict the gender of the subject based on the gender of the cataphoric pronoun. Furthermore, we did not observe an interaction between subject prediction and Binding Condition B in the L2 LM's behavior. This finding stands in contrast to the results reported by van Gompel and Liversedge (2003), where human readers exhibited a preference for linking the pronoun and the subject, and reading times were slowed in the presence of a gender mismatch. Additionally, most L1 LMs displayed an interaction between subject prediction and Binding Condition B, further distinguishing their behavior from that of the L2 LM. These discrepancies in behavior between the L2 LM and its L1 counterparts were particularly evident when comparing the results to those obtained in the second experiment.

4.4 Interaction between Binding Condition B and predictive processing

In the previous experiment, we did not observe a significant interaction that would indicate gender mismatch effects. However, Kush and Dillon (2021) suggest that syntactic predictions are influenced by whether Binding Condition B allows for coreference between the cataphoric pronoun and the subject. This is an important aspect currently under investigation. Despite the L2 LM's failure to accurately predict subject nouns, it is important to gather further evidence regarding whether the L2 LM's predictions are impacted by Binding Condition B. To explore this, we replicated the experiments conducted by Kush and Dillon and compared the results with those obtained from human participants and L1 LSTM LMs. The stimuli used in our study were adapted from Kush and Dillon's work and are presented in (12) and (13).

(12) a. Before spotting *him* at yoga class, the man secretly followed Christian around town.
b. Before spotting *his instructor* at yoga class, the man secretly followed Christian around town.

In (12a), the pronoun *him* cannot co-refer with *the man* because of Binding Condition B. In (12b) and (13), however, coreference between *his* and *the man* is possible. Given this contrast, we employed another version of (12b) in the last experiment:

(13) Before anyone spotted *him* at yoga class, the man secretly followed Christian around town.

Based on the objectives of these experiments, we aimed to investigate whether the gender mismatch effect for cataphora is influenced by Binding Condition B. Specifically, we examined whether there is increased surprisal when the cataphora's gender does not match that of the subsequent subject. In the first experiment of this section, the L2 LM did not exhibit any effect of Binding Condition B, as there was no observed gender mismatch effect. Additionally, we did not find a mismatch effect in cases where there was no Binding Condition B constraint.

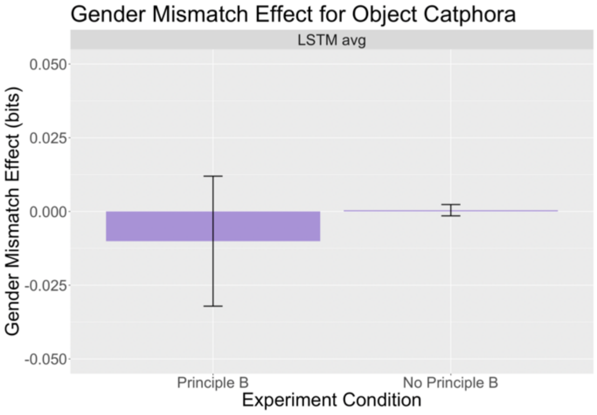

Figure 9. The L2 LM's Gender Mismatch Effect for Subject Following a Cataphoric Object Pronoun (1)

In the last experiment of this study, we used the same pronoun (*him* and *her*) form while modulating the effects of Binding Condition B violations, as in (13). The results are shown in Figure 10.

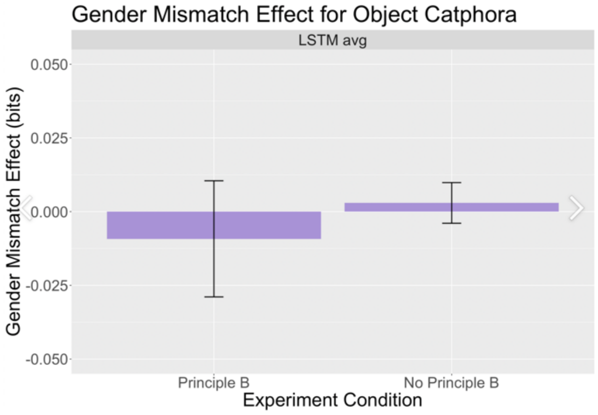

Figure 10. The L2 LM's Gender Mismatch Effect for Subject Following a Cataphoric Object Pronoun (2)

Similarly, the L2 LM did not exhibit any effect of Binding Condition B, indicating a lack of gender mismatch effect. This observation aligns with the findings from the first experiment in this sub-section, where the L2 LM demonstrated consistent behavior across both experimental conditions, regardless of the presence of Principe B violations.

Overall, the L2 LM did not demonstrate the ability to predict the gender of the subject when using object pronouns, even in cases where coreference was grammatically permissible. In contrast to L1 humans and L1 LMs, the L2 LM only acquired knowledge of Binding Condition B in specific contexts. Therefore, our experiments failed to uncover human-like behaviors across all tested conditions.

5. Discussion and Conclusion

To address the two research issues mentioned in Section 2, we have replicated Davis's (2022) experiments to examine the processing behavior of the L2 neural LM in terms of predicting pronouns. In the parallel fashion to Davis, we have evaluated whether the L2 LM can learn human-like syntactic knowledge and control their predictions in specific linguistic environments. Then, we have compared the L2 LM's behaviors to its L1 counterparts.

According to Davis (2022), L1 neural LMs failed to learn the relation between Binding Condition B and other surface configurations. Even when models qualitatively pattern like humans, they missed the estimated effect size by an order of magnitude. In more complex environments, L1 LMs could not constrain neural LMs' predictions. Binding Condition B in humans was a generalized constraint on co-indexation that can be implicated by a number of different surface configurations. However, L1 LMs learned more specific constraints on pronominal agreement.

Likewise, the L2 language model failed to accurately replicate the human-like processing of online coreference and lacked proficiency in exhibiting a comprehensive range of behaviors related to Binding Condition B. In general, the L2 LM consistently exhibited reduced surprisal values when encountering any masculine antecedent in any position. The gender of the embedded subject exerted a greater influence on surprisal values than the gender of the matrix subject. Notably, the gender of the embedded subject primarily affected the surprisal values in instances where ungrammatical positions for coreference were considered. However, similar to the second experiment, the L2 LM managed to adhere to Binding Condition B. When comparing the first and second experiments, it remains uncertain whether the L2 LM truly possesses an understanding of Binding Condition B. To further investigate this behavior, we conducted additional experiments utilizing diverse stimuli. However, these subsequent experiments demonstrated that the L2 LM failed to exhibit a significant correlation between subject prediction and Binding Condition B. Furthermore, the L2 LM did not accurately predict the gender of the subject based on the cataphoric pronoun's gender. In summary, we have determined that capturing both human-like online coreference processing and acquiring a comprehensive understanding of the intricate patterns associated with Binding Condition B proves to be a challenging task for the L2 LM.

Before concluding this paper, it is essential to highlight the noteworthy processing behavior of L2 learners regarding pronouns. Firstly, Kim et al. (2015) observed that adult L2 learners interpreted pronouns in a distinct manner compared to native speakers, encountering challenges when integrating syntactic information with contextual cues. Furthermore, Seo and Shin (2016) reported experimental findings demonstrating that adult L2 learners could interpret pronouns in a manner resembling that of native speakers in both contextual settings by incorporating syntactic binding conditions. However, they faced greater difficulties in interpreting pronouns in contexts that necessitated a higher degree of integration between syntactic and contextual information, particularly compared to reflexives. It is important to acknowledge that comparing the behaviors observed in the L2 LM to the outcomes of these previous studies is challenging due to the utilization of different experimental conditions (e.g., stimuli and methodology). Nonetheless, the behaviors exhibited by L2 learners in pronoun processing warrant consideration to some extent. Both the L2 LM and L2 learners demonstrate evident challenges in accurately predicting the probability of pronouns in more intricate environments, as evidenced by the findings.

It is important to take into consideration previous literature that has reported that neural LMs can capture certain aspects of human-like linguistic processing, although several challenges still remain unresolved. In their study, Hu et al. (2020) emphasized that the size of the dataset used for training the LMs has a more significant impact on the outputs of neural LMs compared to the variety of LM architectures. While we acknowledge the relevance of this point, we believe that the issue of dataset size poses a more significant concern to Hu et al.'s study. Specifically, in this paper, there are disparities in terms of the training dataset size between the two LSTM models. We recognize that there exists a discrepancy in training dataset size between the L1 and L2 LMs. As we employ the same type of model architecture (LSTM language model) as the L1 LMs, the dataset size can influence the performance of the L2 LM. However, the present research does not address the potential impact of this mismatch on the results, leaving this matter unresolved for future investigations.

References

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

Bhattacharya, D. and van Schijndel, M. 2020. Filler-gaps that neural networks fail to generalize. *Proceedings of the 24th Conference on Computational Natural Language Learning*, 486-495.

Chomsky, N. 1981. *Lectures on Government and Binding: The Pisa lectures*. De Gruyter Mouton.

Chow, W. Y., Lewis, S. and Phillips, C. 2014. Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in psychology* 5, 630.

Choi, S. J., Park, M. K. and Kim, E. 2021. How are Korean neural language models 'surprised' layer wisely?. *Journal of Language Sciences* 28(4), 301-317.

Choi, S. J. and Park, M. K. 2022a. An L2 neural language model of adaptation to dative alternation in English. *The Journal of Modern British & American Language & Literature* 40(1), 143-159.

Choi, S. J. and Park, M. K. 2022b. Syntactic priming by L2 LSTM language models. *The Journal of Studies in Language* 22, 547-562.

Choi, S. J. and Park, M. K. 2022c. An L2 neural Language model of adaptation. *Korean Journal of English Language and Linguistics* 37(4), 475-489.

Choi, S. J. and Park, M. K. 2022d. Syntactic priming in the L2 neural language model. *The Journal of Linguistic Science* 103, 81-104.

Clifton, C., Kennison, S. M. and Albrecht, J. E. 1997. Reading the words *her, his, him*: implications for parsing Binding Conditions based on frequency and on structure. *Journal of Memory and language* 36(2), 276-292.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. https://doi.org/10.48550/arXiv.1901.02860

Davis, F. L. 2022. *On the Limitations of Data: Mismatches between Neural Models of Language and Humans*. Doctoral dissertation, Cornell University, Ithaca, NY, USA.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T. and Baroni, M. 2018. Colorless green recurrent networks dream hierarchically. https://doi.org/10.48550/arXiv.1803.11138

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 1-8.

Hu, J., Gauthier, J., Qian, P., Wilcox, E. and Levy, R. P. 2020. A systematic assessment of syntactic generalization in neural language models. https://doi.org/10.48550/arXiv.2005.03692

Jumelet, J. and Hupkes, D. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. https://doi.org/10.48550/arXiv.1808.10627

Kennison, S. M. 2003. Comprehending the pronouns *her*, *him*, and *his*: Implications for theories of referential processing. *Journal of memory and language* 49(3), 335-352.

Kim, E. 2020. The ability of L2 LSTM language models to learn the filler-gap dependency. *Journal of the Korea Society of Computer and Information* 25(11), 27-40.

Kim, E., Montrul, S. and Yoon, J. 2015. The on-line processing of binding Binding Conditions in second language acquisition: Evidence from eye tracking. *Applied Psycholinguistics* 36(6), 1317-1374.

Kush, D. and Dillon, B. 2021. Binding Condition B constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language* 120, 104254.

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

Linzen, T. and Leonard, B. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. https://doi.org/10.48550/arXiv.1807.06882

Liu, X., He, P., Chen, W. and Gao, J. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. https://doi.org/10.48550/arXiv.1904.09482

Marvin, R. and Linzen, T. 2018. Targeted syntactic evaluation of language models. https://doi.org/10.48550/arXiv.1808.09031

Merity, S., Xiong, C., Bradbury, J. and Socher, R. 2016. Pointer sentinel mixture models. https://doi.org/10.48550/arXiv.1609.07843

Nicol, J. L. 1988. *Coreference Processing during Sentence Comprehension*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA.

Nicol, J. and Swinney, D. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of psycholinguistic research* 18, 5-19.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8), 9.

Seo, H. J. and Shin, J. A. 2016. L2 processing of English pronouns and reflexives: Evidence from eye-movements. *Korean Journal of English Language and Linguistics* 16(4), 879-901.

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

Van Gompel, R. P. and Liversedge, S. P. 2003. The influence of morphological information on cataphoric pronoun assignment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(1), 128.

Van Schijndel, M. and Linzen, T. 2018. Modeling garden path effects without explicit hierarchical syntax. *Proceedings of Annual Meeting of the Cognitive Science Society: Changing Minds, CogSci 2018,* 2603-2608.

Van Schijndel, M. and Linzen, T. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science* 45(6), 1-31.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F. and Bowman, S. R. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8, 377-392.

Wilcox, E., Levy, R., Morita, T. and Futrell, R. 2018. What do RNN language models learn about filler-gap dependencies?. https://doi.org/10.48550/arXiv.1809.00042

Wilcox, E., Levy, R. and Futrell, R. 2019. What syntactic structures block dependencies in RNN language models?. https://doi.org/10.48550/arXiv.1905.10431

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary