The Journal of Studies in Language

The Journal of Studies in Language 37(4), 475-489 (2022)

Syntactic Priming by L2 LSTM Language Models

Choi, Sunjoo* and Park, Myung-Kwan**

Dongguk University

***First Author / **Corresponding Author**

ABSTRACT

The Journal of Studies in Language **37.4, 475-489**. Neural(-network) language models (LMs) have recently been successful in performing the tasks that require sensitivity to syntactic structure. We provide further evidence for their sensitivity to syntactic structure by showing that compared to adding a non-adaptive counterpart to it, adding an adaptation-as-priming paradigm to L2 LSTM LMs improves their ability to track abstract structure. By applying a gradient similarity metric between structures, this mechanism allows us to reconstruct the organization of the L2 LMs' syntactic representational space. In so doing, we discover that sentences with a particular type of relative clauses behave in a similar fashion to other sentences with the same type of relative clauses in the L2 LMs' representation space, in keeping with the recent studies of L1 LM adaptation. We also demonstrate that the similarity between given sentences is not affected by specific words in sentences. Our results show that the L2 LMs have the ability to track abstract structural properties of sentences, just as L1 LMs do. **(Dongguk University)**

Keywords: syntactic priming, neural language model, adaptation, L2 LM, representational space

https://doi.org/10.18627/jslg.37.4.202202.475

pISSN : 1225-4770
eISSN : 2671-6151

Received: January 10, 2022

Revised: February 04, 2022

Accepted: February 12, 2022

1. Introduction

Recently, language models (LMs) based on Long Short Term Memory (LSTM) Recurrent Neural Networks (RNNs) have been shown to track abstract structural properties of sentences. Applying a simple adaptation mechanism to such a neural LM, van Schijndel and Linzen (2018) demonstrate that the neural LM adapts not only to lexical items but also to abstract syntactic constructions, just as humans do. They propose a method of continuously adapting a neural LM and testing the method's psycholinguistic plausibility. They show that LM adaptation improves their ability to predict human reading times using a neural LM. Another way of showing this ability to track abstract structural properties of sentences is the syntactic priming paradigm assumed to be part of abstract linguistic knowledge.

- This work was supported by the Dongguk University Research Fund of 2021 (S-2021-G0001-00116).

The study using the syntactic priming paradigm is based on the empirical finding that people have a tendency to repeat the types of sentences that they have recently encountered. For example, a person who has just produced a double object construction (e.g., *Jane gave Michael a toy*) is more likely to produce another double object construction when describing a transfer situation (e.g., *John sent Ben the files*) than to produce a prepositional object construction describing the same event (e.g., *John sent the files to Ben*). Given this, the degree to which one structure primes another allows us to investigate how the representations of sentences are organized. Drawing on the syntactic priming paradigm, Prasad et al. (2019) establish a gradient similarity metric between structures and use this technique to investigate the organization of a neural LM's syntactic representational space. In so doing, Prasad et al. demonstrate that the LSTM RNN (henceforth, LSTM) LM's representations of different types of sentences are organized hierarchically in a linguistically interpretable manner, showing that the LM tracks abstract structural properties of the sentence.

Given an LM's ability to recognize abstract structural properties of sentences, we leverage the previous studies to investigate how L2 LSTM LMs implement the syntactic priming paradigm, compared to L1 LM/human performances. Namely, adopting the methodology employed by the previous works, we examine how much L2 neural LMs can carry out the syntactic priming performances, compared to L1 neural LMs and human L1 speakers.

In general, syntactic priming effects have been studied via psycholinguistic experiments. However, recently neural LMs built on LSTM RNNs have been shown to make adequate syntactic expectations (Gulordava et al., 2018; Linzen et al., 2016) and to make human-like reading time predictions engaging in doing especially syntactic tasks (van Schijndel and Linzen, 2018). Van Schijndel and Linzen (2018) model cumulative priming in RNNs by adapting fully trained neural LMs to new stimuli. They report that when an RNN LM was adapted to a small number of sentences with shared syntactic structure, the surprisal or unexpectancy for new sentences with the same structure decreased. Based on the results, they conclude that the LM's representations of sentences have information about abstract syntactic structure. In a similar vein, Prasad et al. (2019) demonstrate that LSTM LMs' representations of different types of sentences with relative clauses (RCs) are organized in a linguistically interpretable manner. Particularly, sentences with a specific type of RC turn out to be most similar in syntactic priming effects to other sentences with the same type of RC. This result indicates that neural LMs can track abstract structural properties of sentences. Below we are going to review Prasad et al.'s study in more details in Section 3 below.

As mentioned above, building on the previous works on syntactic priming in L1 neural LMs, in this paper we explore a way of adapting a neural LM to linguistic stimuli and use this novel technique to investigate the organization of an LSTM LM's syntactic representational space focusing on L2 neural LMs, at the same time leveraging the syntactic priming paradigm from psycholinguistics. We thus first build L2 LSTM LMs using sentence datasets in the L2 corpus that L2 learners can potentially encounter in their English learning. We then investigate whether L2 LSTM LMs' predictions are consistent with the specific syntactic structures of sentences. To this aim we measure the extent to which L2 neural LMs' predictions for a controlled syntactic structure are affected by their recent experiences with relevant structure. Lastly we compare the adaptation values from L2 LSTM LMs with their counterparts from L1 LSTM LMs.

This paper is organized as follows. In section 2 we briefly introduce the notion of syntactic priming effects as the background of our research. In section 3 we review Prasad et al.'s (2019) study in greater details. Section 4 sets up the experiment configuration. Section 5 investigates how L2 neural LMs represent sentences with relative clauses and coordination structures. Section 6 discusses the results of our experiments. Section 7 wraps with a conclusion.

2. Syntactic Priming in Humans and Neural LMs

Syntactic priming is a speaker's tendency to reuse the same structural pattern as the one that was previously encountered (Bock, 1986). For example, dative events can be expressed using two roughly equivalent English constructions, as in (2). Work in psycholinguistics has shown that the recent experience with one of these variants increases the probability of producing that variant. In other words, after encountering examples like (1) (which is called the prime sentence), people expect to use examples like (2a), which shares syntactic structure with the prime sentence, rather than (2b). Simply speaking, (1) shares the syntactic structure (that is, VP -> V NP NP) with (2a), but not with (2b).

(1) The boy threw the dog the ball.

(2) a. The chef made the guest some excellent pizza.
b. The chef made an excellent pizza for the guest.

The syntactic priming paradigm has been applied to investigate whether the representations of given sentences have shared structure. If (1) primes (2a) rather than (2b), we can infer that the representation of (1) is more similar in structural property to that of (2a) than to that of (2b).

Notably, priming effects can be cumulative. It means that priming effects are accumulative across many utterances. In this environment, sentences with a shared structure Sx become easier to process when preceded by sentences with the same structure Sx than when preceded by *n* sentences with the different structure Sy (Kaschak et al., 2011). The cumulative priming effects indicate that people begin to expect a structure with a greater probability with increased exposure to that structure (Chang et al., 2006). Cumulative priming enables us to investigate how sentences are relevant in the human representation space, in the different way that non-cumulative priming does. When people are exposed to sentences with a shared structure Sx, if there is a decrease in surprisal when they are tested on other sentences with the same structure Sx than other sentences with the different structure Sy, we can infer that the representations of sentences with Sx are more similar in structural properties to each other than to the representations of sentences with Sy.

We can extend this line of research to examine whether, like human expectations, a neural LM's expectations are consistent with its experience with the syntactic structures of sentences. To confirm this, we can calculate the extent to which a neural LM's expectation for a specific syntactic structure is influenced by the recent experiences with relevant structures. van Schijndel and Linzen (2018) in fact model cumulative priming to investigate an L1 neural LM's ability to predict human reading times. They address the following two research questions. First, when a neural LM assigns a higher probability to the words that have recently occurred/been experienced, how much is the former affected by the latter (Kuhn and de Mori, 1990)? Second, how much of the improvement is due to the neural LM's adaptation to syntactic representations (Dubey et al., 2006)? Answering these questions, they take a fully trained LSTM RNN LM and continue to train it on a small set of sentences. They show that LM adaptation significantly improves their ability to predict human reading times using the neural LM. The experiments with controlled materials show that the neural LM adapts to both specific vocabulary items and abstract syntactic constructions. The adaptation mechanism significantly improves the LSTM RNN LM's word prediction accuracy, as found for humans. Next, we review in some details Prasad et al.'s (2019) study in keeping with van Schijndel and Linzen's study, as it serves as a basis for the subsequent discussion in Section 4.

3. Prasad et al. (2019)

Prasad et al. (2019) also prime a fully trained model with a structure by adapting it to a small number of sentences having that structure (van Schijndel and Linzen, 2018). Then, they measure the change in surprisal after adaptation when a neural LM is tested either on sentences with the same structure or sentences with different but related structures. They adopt a similarity metric between sentences with Sx and those with Sy in the neural LM's representation space by adapting the neural LM to sentences with Sx. Then, they calculate the change in surprisal for sentences with Sy. This value shows to what extent sentences with Sx prime sentences with Sy. If Sx and Sy are similar in structural property to each other in the LM's representation space, then A (Y | X)[1] is higher than 0. It means that sentences with Sx makes the LM give a higher probability to sentences with Sy. However, if Sx and Sy are different in structural property to each other, then A (Y | X) renders 0. This is because sentences with Sx do not change a probability for sentences with Sy.

For calculating the measure of similarity, they apply the method to different types of sentences with relative clauses (RCs). They test five types of RCs as follows. First, in an active RC, the gap is in the subject position of the relative clause.

(3) My friend that _ liked the movie ...

Second, in a passive subject RC, the gap is in the subject position of the relative clause, and the RC verb is passive. Passive RCs can be reduced or unreduced as shown in (4).

(4) a. The movie that _ was liked by my friend ...
b. The movie _ liked by my friend ...

Third, an object RC has a gap in the object position of the relative clause. Object RCs can also be reduced or unreduced, as in (5).

(5) a. The movie that my friend liked ...
b. The movie my friend liked ...

Finally, they also used two more conditions with verb coordination: one with almost the same word order and lexical content as active RCs, and another with almost the same word order and lexical content as passive RCs and object RCs[2], as illustrated in (6) and (7).

(6) My friend liked the movie and ...

(7) The movie liked my friend and ...

They employed these templates to generate five experimental lists, and each list included a pair of adaptation and test

1) The notation A (Y | X) refers to the change in surprisal.
2) To maintain the same word order as object and passive RCs, the subject of the coordinated verb phrases is an NP. For this reason, some of the sentences in this condition are implausible (Prasad et al., 2019: 3).

sets. Each test set contained 50 sentences, and each adaptation set contained 20 sentences.

Applying the dataset, they built 75 versions of the LSTM language models trained by van Schijndel et al. (2019); these LMs varied in the number of hidden units per layer (100, 200, 400, 800, 1600) and the number of tokens they were trained on (2 million, 10 million, or 20 million). As in van Schijndel and Linzen (2018), they trained neural LMs on five disjoint subsets of the WikiText-103 corpus.

For all the different types of structure, they computed the similarity between one structure and other structures mentioned above, as illustrated in Figure 1. The surprisal values were averaged across the entire sentence.[3)]

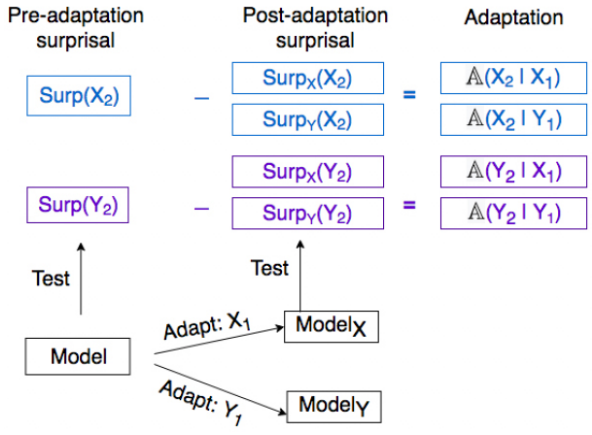

Fig. 1. A schema for calculating the adaptation effects (cited from Prasad et al., 2019: 4)

Given the process described above, they expected sentences that shared the same specific structure to behave in a more similar fashion to each other than just lexically matched sentences that did not share the structure. In result, this prediction was confirmed for all of seven structures, as in Figure 2a. In addition, as in Figure 2b, the adaptation effects for the models adapted to RCs were greater when they were tested on sentences with other types of RCs than when they were tested on sentences with coordination. Likewise, the adaptation effects for the neural LMs adapted to one type of coordination were statistically significant when the models were tested on sentences with the other type of coordination.

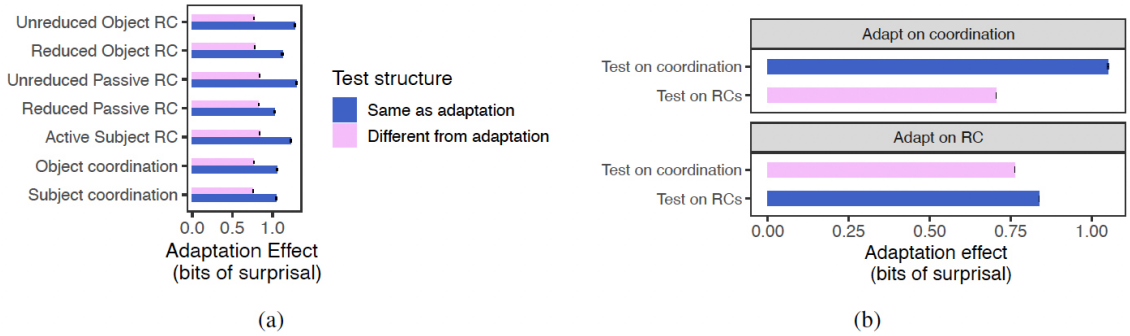

Fig. 2. Adaptation effects averaged across all 75 models (cited from Prasad et al., 2019: 5)

3) Unknown words are not included.

Furthermore, they tested the different types of RCs based on the two linguistically interpretable features (that is, reduction and passivity). This is because they investigated whether the neural LMs could track either, both or none of the two features at issue. The adaptation effects were statistically significant when there was a match in both features. They finally concluded that the neural LMs can track both of these features, as illustrated in Figure 3.

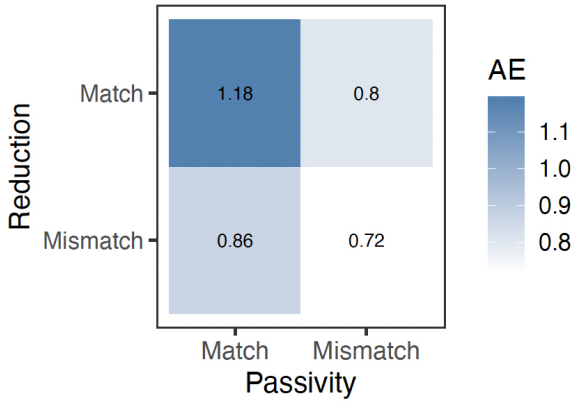

Fig. 3. The adaptation effects when the models adapted to sentences with reduced and unreduced RCs were tested on sentences that matched only in reduction, matched only in passivity, matched in both reduction and passivity, or sentences that matched in neither. (cited from Prasad et al., 2019: 6)

Lastly, to confirm whether the similarity between the members of the sentence classes/types was driven by the presence of shared function words, Prasad et al. compared the representation space of the trained modes with that of the models trained on no data (henceforth, the baseline models). To explore this issue, they measured the distance between the members and the non-members of the three interpretable sentence classes/types: (i) sentences which have the same type of RC, (ii) sentences that match in reduction, or (iii) sentences that have any type of RC. As shown in [4a] below, for all the three sentence classes, the sentences that belonged to one of these classes behaved more similarly to each other than those that did not belong to that class. This result is somewhat unexpected because the sentences at hand do not have any function word that is shared by all the RC-class sentences.

Besides, they examined agreement prediction accuracy on object RCs. This is because the previous work of van Schijndel and Linzen (2018) reports that neural LMs have a difficulty in predicting the number feature of main verbs if the main clause subject NP is modified by an object RC. Prasad et al. made two assumptions: when neural LMs take object RCs not to be in the same class with other RCs, there are few training examples from which the models can learn about agreement prediction that the subject NP is modified by an object RC. Second, when neural LMs take object RCs to be in the same class with other types of RCs, they can generalize from training sentences of subject-verb agreement that the subject NP is modified by other RCs. Given that, they found that there was an increase in accuracy as the number of hidden units increased, as in [4b]. Notably, the similarity between object RCs and other types of RCs turned out to have nothing to do with agreement prediction. Eventually, they did not find any evidence supporting the two assumptions.

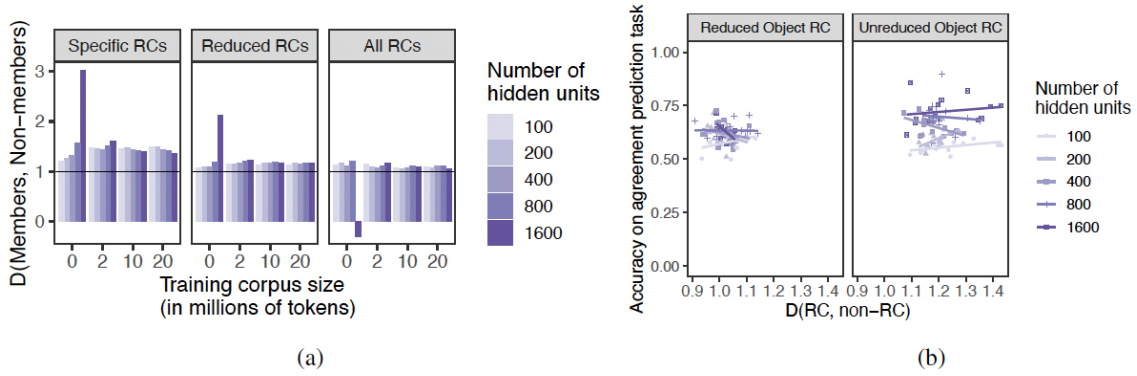

Fig. 4. (a) Effect of hidden layer size and corpus size (b) Agreement prediction accuracy on reduced object RCs and unreduced object RCs (cited from Prasad et al., 2019: 8)

Taking stock, by applying the syntactic priming paradigm, Prasad et al. successfully reconstruct the organization of L1 neural LMs' representational space via a gradient similarity metric. In so doing, they find that L1 neural LMs can track abstract structural properties of the sentences tested. Building on the results, we turn to investigate how L2 LSTM LMs implement the syntactic priming paradigm, compared to their L1 counterpart models. To carry out this aim, we build L2 LSTM LMs trained on the L2 English dataset. Drawing on the adaptation-as-priming paradigm, we gain insight into the representations of sentences with relative clauses in the L2 LSTM LMs. The follow-up experiments show that the L2 LSTM LMs can adapt to abstract syntactic constructions as well as specific vocabulary items, as L1 neural LMs do. We now describe how Prasad et al.'s study is replicated for the L2 LSTM LMs.

4. Methods

4.1 LSTM Language Models for L2ers

The model which we adopt for the present study is the Gulordava LM (Gulordava et al., 2018); it is chosen for its previous success in learning the subject-verb number agreement task. The Gulordava LM is pre-trained on 90 million tokens of English Wikipedia and has two hidden layers of 650 units each. Adopting the Gulordava LM architecture, we design 9 of the LSTM language models employing the model architecture proposed by Gulordava et al. (2018); these L2 LSTM LMs vary in the number of hidden units per layer (100, 200, 400) and the number of tokens they are trained on (7 million, 10 million, 13 million). The L2 LMs are implemented using the L2 leaner English dataset collected from English textbooks for Korean L2ers based on the 11 middle-school and 12 highschool English textbooks published in Korea in 2001, and the 19 middle-school and 12 highschool English textbooks published in Korean in 2009, as well as EBS-CSAT English Prep Books published in 2016-2018.

4.2 Experimental Environment

We follow the same factorial approach and adopt the code of Prasad et al. (2019). In addition, all the experimental items are taken from Prasad et al (*ibid.*). Each list has a pair of adaptation and test sets with minimal lexical overlap between them. Each adaptation set contains 20 sentences, and each test set includes 50, as described in Table 1.

Table 1. Examples of the sentences involving seven types of structure

Structure	Example
Unreduced Object	The teachers that my shy partner admired last year moved the stone remorsefully.
Reduced Object	The teachers my shy partner admired last year moved the stone remorsefully.
Unreduced Passive	The teachers that were admired by my shy partner last year moved the stone remorsefully.
Reduced Passive	The teachers admired by my shy partner last year moved the stone remorsefully.
Active Subject	My shy partner that admired the teachers last year stirred the water.
Subject Coordination	My shy partner admired the teachers last year and stirred the water.
Object Coordination	The teachers admired my shy partner last year and moved the stone remorsefully.

4.3 Measuring the Adaptation Effect

The general process for the experiment is represented in Figure 1. We define a similarity metric between structures by adapting the L2 neural LMs to structure Sx. We then calculate the change in surprisal[4] for sentences with the same structure (that is, Sy). Simply put, we measure to some extent that the sentences with structure Sx prime sentences with Sy. If Sx and Sy are structurally similar in the L2 neural LMs' representation space, we expect the adaptation effects to be a positive number. On the other hand, if they are structurally different in the L2 neural LMs' representation space, the adaptation effects will be zero because there is no probability for sentences with Sy.

5. Results

We used the linear mixed effects models (Pinheiro and Bates, 2000) to test for statistical significance[5]. The L2 neural LMs predict sentences that share the same structure to behave in a more similar fashion than just lexically matched sentences that do not share the structure, as L1 neural LMs did. The results show that this prediction was borne out for the seven structures we employed, as described in Figure 5. The graph shows how the L2 neural LMs were adapted to each of the structures and then tested on the same structure (as represented by the bottom blue bars), or different structure (as represented by top pink bars).

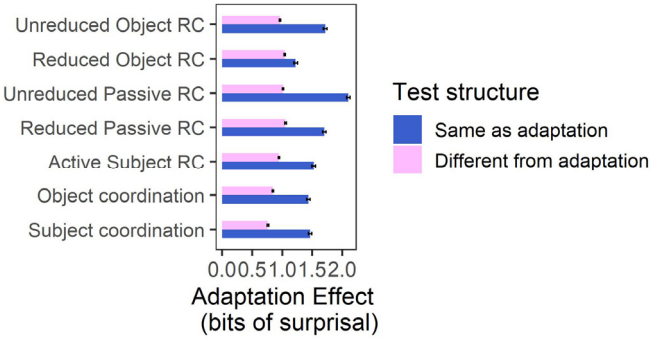

Fig. 5. Adaptation effect for seven target structures[6] [7]

4) The surprisal values were averaged across the entire sentences.
5) We report highly significant results only.

As mentioned in Section 3, we also employed two coordination conditions that were structurally identical but different in semantic plausibility. Most of the examples of object coordination condition are somewhat semantically implausible; however, the examples of subject coordination condition are semantically plausible (see footnote 2). If structurally identical sentences are close together regardless of semantic plausibility, we predicted sentences with coordination to behave in a more similar fashion to each other than lexically matched sentences with RCs. As predicted, the adaptation effects for the L2 neural LMs adapted to one type of coordination were statistically significant when the L2 neural LMs were tested on the sentences with the other type of coordination. These effects were greater than when they were tested on sentences with RCs, as in the top panel of Figure 6.

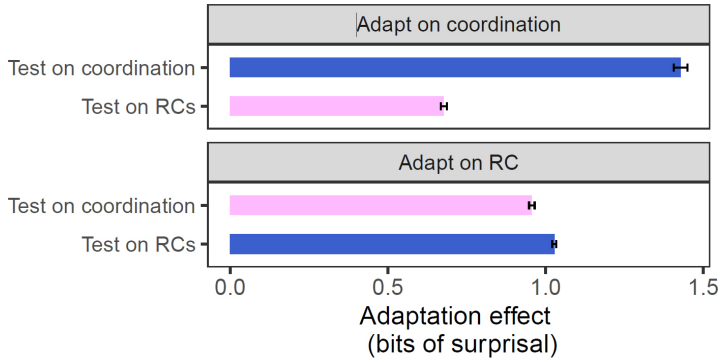

Fig. 6. Similarity between sentences with different types of coordination and RCs

The noteworthy point is that sentences with different types of RCs have a common property at an abstract level: they have a gap. They differ in this respect from sentences with coordination. If the L2 neural LMs keep track of whether or not a sentence contains a gap, we predicted sentences with different types of RCs behave in a more similar fashion in the L2 neural LMs' representation space than lexically identical sentences without a gap. This prediction was fulfilled, as in the bottom panel of Figure 6. The adaptation effects were statistically significant when they were tested on the sentences with other types of RCs. These effect were greater than when they were tested on the sentences with coordination. These observed patterns may not be so surprising. The results indicate that the L2 neural LMs can recognize whether or not a sentence contains a gap, regardless of the lexical overlap between prime and target sentences.

Furthermore, based on linguistically interpretable features, the different classes of RCs can be classified according to reduction and passivity. To wit, reduction is divided into reduced passive and object RCs distinguished from unreduced passive and object RCs . In the same way, passivity is divided into reduced and unreduced passive RCs distinguished from reduced and unreduced object RCs. Given these classifications, we further tested whether the L2 neural LMs can detect these two linguistic features by comparing the similarity between sentences that share one feature but not the other, with the similarity between sentences that share neither feature. We predicted that the adaptation effects would be more significant when there was a match in one feature than when there was no match in any of the features. When the

6) The adaptation effects were averaged across all the nine models when they were adapted to each of the structures and tested on either same or different structure.

7) The results reported in Figure 5, Figure 6, and Figure 8 are taken from Kim (2022), who performed an NLP-based study of the same issues as explored in our linguistics-based study.

prediction is achieved, we can naturally infer that the L2 neural LMs can track whether or not sentences have linguistic features. As a result, we discovered that the L2 neural LMs were able to recognize the two features mentioned before, as illustrated in Figure 7.

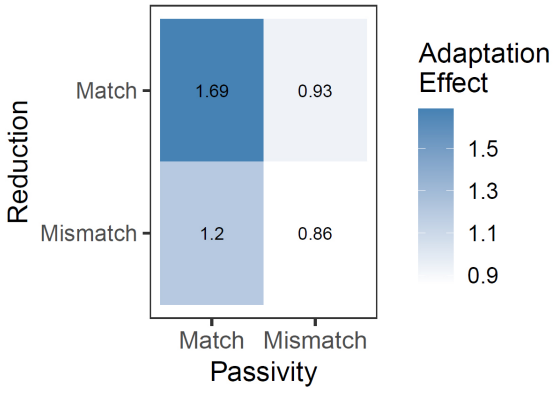

Fig. 7. The adaptation effects between sentences with different RCs[8)]

As in Figure 7, when the adaptation and test sentences matched only in passivity, the adaptation effects were slightly significant than when the adaptation and test sentences matched only in reduction. It means that the L2 neural LMs recognized that the examples like (8) were more similar to (9) than to (10). In other words, passivity affected the similarity between the sentences more significantly than reduction.

(8) The guardian my friend admired fried the chicken.

(9) The guardian that my friend admired fried the chicken.

(10) The guardian admired by my friend fried the chicken.

We have so far found that sentences that share linguistic features behave in a more similar fashion to each other in the L2 neural LMs' representation space. The question raised here is what properties of sentences influenced the similarity between members of the three classes. As mentioned above, most of the sentences shared function words. For this reason, we assumed that the similarity might be affected by the presence of function words. To evaluate the validity of this assumption, we compared the representation space of the L2 neural LMs we tested in the previous section (that is, the trained models) with the representation space of the L2 neural LMs trained without data (that is, the baseline models). As in the baseline models, there was no lexical overlap in content words between adaptation and test sets. In this environment, we predicted that the similarity between sentences would be affected by function words. In so doing, if the similarity between the sentences in the representation space of the trained models was driven by other factors, we would expect this similarity to be greater than the similarity between these sentences in the representation space of the baseline models.

8) The results indicate when the L2 LMs adapted to examples with reduced/unreduced RCs are tested on examples that match only in reduction (top right), on those that match only in passivity (bottom right), on those that match in both reduction and passivity (top left), or examples that match in neither (bottom right).

Then, we measured a distance between the sentences that belonged to a group Sx and sentences that did not belong to a group Sx, as schematized below. This way of comparison was made because we cannot simply measure adaptation effects to compare the similarity between the sentences in the representation spaces of both models.

$$\mathbb{D}(S_X, \neg S_X) = \frac{AE(X_2 \mid X_1)}{AE(\neg X_2 \mid X_1)}$$

We assumed that the distance value would be greater than one if the sentences that belonged to the same group were more similar to each other than the sentences that did not belong to the group. In this respect, we examined the distance between members and non-members for the three linguistically interpretable groups. The first group includes the sentences having the same type of RC, and the second group includes the sentences that match in their reduction. The third group includes the sentences that have any type of RC. According to our baseline models, for all the three groups of sentences, those that belonged to one of these groups behaved in a more similar fashion to each other than sentences that did not belong to that group, as shown in Figure 8. As noted by Prasad et al., this result is instructive since there is no common function word within all the RC conditions.

We also found that when the sentences shared some function words, for the trained models the distance between the sentences that belonged to one group and the sentences that did not was significant than for the baseline models. This means that the similarity between different groups of sentences within the trained models was affected by other factors than function words. Besides, differently from L1 neural LMs, in L2 neural LMs the distance between members and non-members was not affected by the number of training tokens.

Furthermore, we discovered that when the sentences did not share any function words, in L2 neural LMs the distance between sentences that belonged to one group and sentences that did not belong to that group did not differ between the trained models and the baseline models. Simply put, the similarity between the sentences in the representation space of trained models was affected by the presence/absence of lexical items.

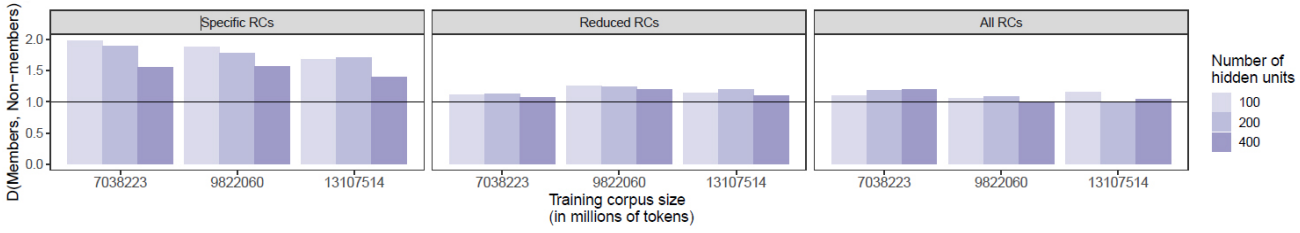

Fig. 8. Effect of model size and training corpus size on the distance between sentences

Marvin and Linzen (2018) reported a dataset that was intended to probe the grammaticality predictions of L1 LMs. Given the dataset, they found that the models had a difficulty in predicting the number feature of the main verb when the main clause subject was modified by an object RCs. Surprisingly, the models performed well if the main clause was modified by an active subject RC. On the other hand, the models performed poorly in predicting that (11a) with an object RC should have a higher probability of being grammatical than (11b). But they were better at predicting that (12a) with a subject RC should have a higher probability of being grammatical than (12b).

(11) Agreement in an object relative clause
a. The farmer that the parents love swims.
b. *The farmer that the parents love swim.

(12) Agreement in a subject relative clause
a. The farmer that loves the parents swims.
b. *The farmer that loves the parents swim.

(Marvin and Linzen, 2018: 10-11)

Roland et al. (2007) explained the observed contrast between object and subject RCs, relying on the fact that object RCs are pretty infrequent in general. As mentioned above, we also tested Prasad et al.'s hypothesis that agreement prediction on object RCs would be higher in L2 neural LMs when the representation space of object RCs behaves in a more similar fashion to the representation of other RCs. To wit, they suggested the following hypothesis: if a neural LM takes object RCs to be unrelated to other RCs, there are likely to be few training examples from which the neural LM can learn about subject-verb agreement that the subject NP is modified by an object RC. If the neural LM takes object RCs to be related to other RCs, they could learn to generalize from training sentences of subject-verb agreement that the subject NP is modified by other RCs than an object RC.

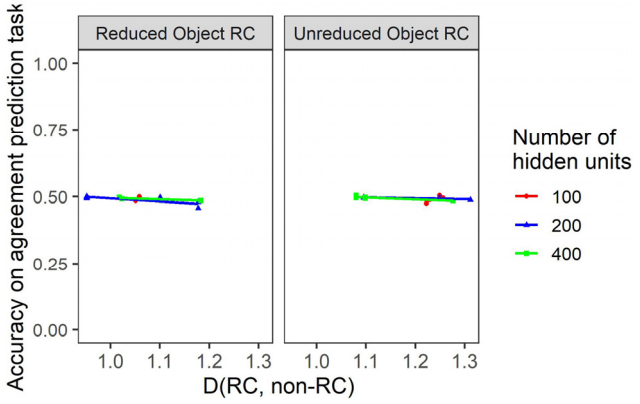

Fig. 9. Agreement prediction accuracy on reduced object RCs and unreduced object RCs

As reported in the previous section, the results from L1 neural LMs showed that there was an increase in accuracy as the number of hidden units increased. Differently from L1 neural LMs, for L2 neural LMs there was a mere change in accuracy, as shown in Figure 9. The similarity between the two conditions was not significantly relevant to agreement prediction. Thus, we failed to find any evidence for fulfilling Prasad et al.'s hypothesis at issue.

Taken together, these results from the L2 LSTM LMs have demonstrated that they were able to track abstract structural properties of sentences. We have shown that when the L2 LSTM LMs were adapted to sentences with a shared syntactic structure, the similarity between adaptation and test structures was statistically significant. In other words, the sentences that belonged to the same group of sentences behaved in a more similar fashion to each other in the representation space of the L2 LSTM LMs.

6. Discussion

Based on the syntactic priming paradigm, we adopted a new method of examining how the representations of sentences in the L2 LSTM LMs are organized. Based on relative clauses constructions in English, we have found that the representations are organized in a linguistically interpretable manner. We have also probed whether this observed pattern was driven by function words that were shared among sentences. We have discovered that the sentences that belonged to the same group were more similar to each other in the representation space of the trained L2 LSTM LMs than the baseline LMs that were not trained on any data, as in that of the L1 counterparts. As noted by Prasad et al., the results indicate that the trained L2 LSM LMs were able to track abstract structural properties of the sentence. Besides, when sentences involved a gap (having a non-lexically observable property), those sentences were similar to each other in the representation space of the L2 LSTM LMs.

Furthermore, we have examined the LMs' accuracy on number agreement. We focused on the poor performance of the LMs when the main clause subject NP was modified by object RCs. Unlike the L1 LMs, for the L2 LSTM LMs there was no difference in accuracy as the number of hidden units increased. We have also tested the hypothesis that accuracy on agreement with object RCs would increase as the similarity between object RCs and other types of RCs increased. We conclude that the similarity between object RCs and other types of RCs is not consistent with agreement predictions.

Overall, the L2 LSTM models are similar in observed patterns to the L1 counterparts except agreement prediction accuracy. Before leaving this section, it is to be mentioned that Prasad et al. (2019) point out that the sentences belonging to linguistically interpretable classes were more similar to each other in the representation spaces of the models trained on 2 million tokens than in the representation spaces of the models trained on 20 million tokens. They interpret this pattern to suggest that an LM's ability to track abstract structural properties of sentences decreases with an increase in the training corpus size. Though our training data size is not greater than Prasad et al.'s study, our results of the LMs' syntactic priming also do not correlate significantly with the training corpus size. In other words, we do not observe any difference between training corpus size and LMs' performance in syntactic priming.

7. Conclusion

In this study, given the syntactic priming paradigm, we have explored how the representations of sentences with relative clauses are organized within L2 LMs and compared them with L1 models in light of observed patterns. The degree to which one structure primes another has enabled us to measure a similarity metric between the L2 models' representations of structures. We have applied the similarity metric between target sentences (RCs) and thereby have reconstructed the L2 neural LMs's syntactic representational space in the same way with Prasad et al.'s (2019) work. We have discovered that as in the L1 neural LMs, for the L2 neural LMs the different types of sentences with RCs are organized in a linguistically interpretable way. It has been found that the sentences with a specific type of RC are similar to other sentences with the same type of RC in the L2 LMs. Sentences with different types of RCs are, on the other hand, more similar to each other than sentences without RCs. Importantly, we have shown that the similarity is not affected by particular words in sentences. Our results support the previous report that neural LMs are capable of tracking abstract

structural properties of sentences. However, the ability to track abstract structural properties of sentences is not influenced by training corpus size.

In the future, we go further to compare L2 neural LMs' representations with human representations based on the syntactic priming paradigm. In so doing, we can eventually investigate how human-like the LM representations are. We leave the discussion of this and other potential issues for future work.

References

Bock, J. K. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18, 355-387.

Chang, F., G. S. Dell, and K. Bock. 2006. Becoming syntactic. *Psychological Review* 113, 234-272.

Dubey, A., F. Keller, and P. Sturt. 2006. Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. *Proceedings of the 21st international Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 417-424.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1195-1205.

Kaschak, M. P., T. J. Kutta, and C. Schatschneider. 2011. Long-term cumulative structural priming persists for (at least) one week. *Memory and Cognition* 39, 381-388.

Kim, E. 2022. Probing sentence embeddings in L2 learners' LSTM neural language models using adaptation learning. Ms., Shinhan University.

Kuhn, R. and R. De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 570-583.

Linzen, T., E. Dupoux, and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535.

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

Pinheiro, J. and D. Bates. 2000. *Mixed-effects Models in S and S-PLUS*. Springer Science & Business Media.

Prasad, G., M. Van Schijndel, and T. Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. *Proceedings of the 23rd Conference on Computational Natural Language Learning*, 66-76.

Roland, D., F. Dick, and J. L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57, 348-379.

Van Schijndel, M. and T. Linzen. 2018. A neural model of adaptation in reading. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4704-4710.

Van Schijndel, M., A. Mueller, and T. Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5831-5837.

Choi, Sunjoo, Postdoctoral Fellow
30, Pildong-ro 1-gil, Jung-gu, Seoul, 04620, Republic of Korea
Department English, Dongguk University
E-mail: sunjoo@dongguk.edu

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