Korean Journal of English Language and Linguistics, Vol 22, May 2022, pp. 547-562
DOI: 10.15738/kjell.22..202205.547

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

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

http://journal.kasell.or.kr

An L2 Neural Language Model of Adaptation[*]

Sunjoo Choi (Dongguk University) **Myung-Kwan Park** (Dongguk University)

Received: May 9, 2022
Revised: June 18, 2022
Accepted: June 30, 2022

Sunjoo Choi (first author)
Post-Doctor, Division of English Language and Literature, Dongguk University
E-mail: sunjoo3008@gmail.com

Myung-Kwan Park (corresponding author)
Professor, Division of English Language and Literature, Dongguk University
E-mail: korgen2003@naver.com

ABSTRACT

Choi, Sunjoo and Myung-Kwan Park. 2022. An L2 neural language model of adaptation. *Korean Journal of English Language and Linguistics* 22, 547-562.

In recent years, the increasing capacities of neural language models (NLMs) have led to a surge in research into their representations of syntactic structures. A wide range of methods have been used to address the linguistic knowledge that NLMs acquire. In the present study, using the syntactic priming paradigm, we explore the extent to which the L2 LSTM NLM is susceptible to syntactic priming, the phenomenon where the syntactic structure of a sentence makes the same structure more probable in a follow-up sentence. In line with the previous work by van Schijndel and Linzen (2018), we provide further evidence for the issue concerned by showing that the L2 LM adapts to abstract syntactic properties of sentences as well as to lexical items. At the same time we report that the addition of a simple adaptation method to the L2 LSTM NLM does not always improve on the NLM's predictions of human reading times, compared to its non-adaptive counterpart.

KEYWORDS

syntactic priming, adaptation, neural language model, surprisal, perplexity, learning rate

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

1. Introduction

It has been noted that neural(-network) language models (NLMs) are able to represent and learn a broad range of linguistic phenomena (Gulordava et al. 2018, Tenney et al. 2019, Warstadt et al. 2020). However, it remains unknown whether these NLMs indeed encode an abstract notion of syntax in their representational space and to which extent they can acquire specific linguistic constructions. Many previous works have addressed these issues taking inspiration from psycholinguistics, particularly the syntactic priming paradigm (Bhattacharya and van Schijndel 2020, Fine and Jaeger 2016, Prasad et al. 2019, van Schijndel and Linzen 2018, Sinclair et al. 2021). The syntactic priming paradigm in question refers to a tendency of people to reuse the structural pattern that people have experienced when they produce sentences (Bock 1986). For example, when a sentence like *The professor gave his student the paper* is encountered, the same structure is more likely to be used in a subsequent sentence like *The boy passed his dad a letter* than the alternate structure *The boy passed a letter to his dad*. Evidence from syntactic priming provides us with an opportunity to estimate how sentences are similar to each other in humans' (or NLMs) representation space. This paradigm also enables us to gain insight into how human-like the NLM representations are and vice versa.

Earlier work on syntactic priming in Long Short-Term Memory (LSTM) LMs by van Schijndel and Linzen (2018) primed a fully trained model like LSTM with a structure by adapting it to a small number of sentences containing that structure. Then, they demonstrated that the addition of a simple adaptation mechanism to a NLM improves predictions compared to a non-adaptive model. They noted that NLMs adapt abstract syntactic structures as well as lexical items. Then, Prasad et al. (2019) used van Schijndel and Linzen's (*ibid.*) syntactic priming paradigm to explore how recurrent neural network LMs represent sentences with relative clauses (RCs). Prasad et al. measured the change in surprisal values after adaptation when the NLMs are tested either on sentences with same structure or sentences with different but related structures. They argued that the representations of sentences with relative clause are organized in a linguistically interpretable manner. In other words, sentences with a particular type of RC are more similar to other sentences with the same type of RC. In so doing, they supported the claim that the NLMs can learn abstract structural properties of sentences. This line of analysis for NLMs based on previous behavioral findings indicates that the syntactic priming paradigm is a highly useful and effective method for gaining insights into the capacities of NLMs.

Given the recent success in the L1 LM's adaptation as well as the priming behaviors of humans, this research aims to investigate the syntactic priming patterns of the L2 LSTM LM trained on training data collected from English textbooks published in Korea. In other words, adopting the Gulordava model architecture, which was pre-trained on about 90 million tokens of English Wikipedia (i.e., the L1 LSTM LM), we implemented the L2 LSTM LM for our experiments. Based on the L1 LM's ability to prime abstract structural properties of sentences, we leverage the previous work to apply an elaborate experimental method to assess the syntactic priming behaviors of the L2 LSTM LM. For our experiments, we implement the L2 LSTM LM trained on the L2 corpus that Korean L2 English learners can potentially encounter during their English learning in middle and high-school days. Then we adapt the L2 LSTM LM to some syntactic constructions and compare its performance with the counterpart in the non-adaptive version of the model. Conducting a series of computational experiments using the L2 adapted LM, we can examine whether L2 adapted LM's expectations are consistent with the target items. Our findings will show that the L2 adapted LM display inconsistent behaviors regarding priming effects in the subsequent experiments, unlike the L1 adapted LM.

The structure of the paper is as follows. Section 2 touches on syntactic priming effects as the theoretical background of our experiments. Section 3 introduces van Schijndel and Linzen's (2018) study in greater details.

Section 4 introduces the implementation of the L2 LSTM LM. Section 5 reports the results of the L2 adopted LM's performances, concentrating on its abilities to capture priming effects. Section 6 discusses our finding. Section 7 concludes the paper.

2. Syntactic Priming Effects in Humans and Language Models

Syntactic priming has been one of the dominant paradigms in psycholinguistics in investigating whether the representations of two sentences have shared structure. For example, (1) shares the structure with (2a), that is, VP → V NP NP.

(1) The teacher assigned the class the homework.
(2) a. The salesperson sold the customer the product.
b. The salesperson sold the product to the customer.

If (1) primes (2a) more consistently than it primes (2b), we can suspect that the representation of (1) is more similar structurally to that of (2a) than that of (2b). We assume that as a prime structure is processed, it becomes activated and therefore persists across utterances with a range of different syntactic structures (Bock 1986, Bock and Griffin 2000, Pickering and Branigan 1998, Pickering and Ferreira 2008). The rationale behind this behavior is that if speakers have a tendency to reuse aspects of sentence structure, then it means that such structural information is an integral part of the representations built during sentence processing. In tandem, semantic relatedness between prime and target sentences provides boosting effects on syntactic priming (Mahowald et al. 2016).

At the same time, the effects of syntactic priming can be cumulative. Sentences with a shared structure S*x* become progressively easier to process when preceded by *n* sentences with the same structure S*x* than when preceded by n sentences with a different structure S*y* (Kaschak et al. 2011). Simply put, cumulative exposure to prime structure will boost priming effects. Cumulative priming allows us to study how sentences are similar to each other in the human or NLM representation space. It means that when participants encounter sentences with structure S*x*, if there is a greater decrease in surprisal (i.e., unexpectedness) when they are tested on other sentences with S*x* than when they are tested on other sentences with S*y*, we can conjecture that the representations of sentences with S*x* are more related to each other than sentences with S*y*.

van Schijndel and Linzen (2018) reported that when a RNN LM was adapted to a small number of sentences with a shared syntactic structure, the surprisal for new sentences with the same structure decreased. The results demonstrate that the NLM's representations of sentences accommodate the test structure on the basis of prior experience. Recently, Sinclair et al. (2021) also investigated how priming can be used to study the nature of the syntactic knowledge acquired by NLMs. Conducting several experiments, they found strong priming effects when priming was trained with multiple sentences. The priming effects increased as the number of syntactically congruent prime sentences were accumulated. They also noted that when presenting multiple primes with the same syntactic structure, the proximity of such primes had a positive effect on priming.

In this paper, we are to capitalize on the benefits of the cumulative priming paradigm to study the priming behaviors of the L2 LSTM LM and examine its ability to encode abstract structural information. In what follows, we review van Schijndel and Linzen's (2018) study in greater details, as it serves as a basis for our study of the L2 LSTM LM reported in Section 4 and 5.

3. Van Schijndel and Linzen (2018)

3.1 Method

van Schijndel and Linzen (2018) employed a simple method in adapting an NLM as follows: at the end of each new test sentence, the parameters of the NLM were updated based on its cross-entropy loss when predicting that sentence; the new weights are then used to predict the subsequent text sentence. They followed Gulordava LM (Gulordava et al. 2018) as baseline LM which was trained on 90 million English words from Wikipedia. Given this, they conducted three experiments to measure the adaptation effects. In addition, they tested the model on the Natural Stories Corpus (Futrell et al. 2018), which has 10 narratives with self-paced reading times from 181 native English speakers: fairy tales (seven texts) and documentary accounts (three texts).

3.2 L1 Experiment 1: Linguistic Accuracy

First, van Schijndel and Linzen measured how well the adaptive model expected upcoming words using the model's perplexity[1]. To do so, they adapted the model to the first *x* sentence in either of the two types of texts such as documentary texts and fairy tales and then tested it on the *x* + 1 sentence for all *x'*s. As a result, this procedure of adaptation improved on the test perplexity of the adaptive mode, compared to the counterpart of the non-adaptive model (86.99 vs. 141.49).

Crucially, they adapted the model to each genre separately. If the model adapts to stylistic or syntactic patterns in one type of text, adaptation effects are more robust in this type of text than in the other type of text. This prediction is fulfilled, as shown in Table 1. The documentary texts benefited less from adaptation than the fairy tales.

Table 1. L1 LM's Test Perplexity

	Documentary texts	Fairy tales
Non-adaptive surprisal	99.33	160.05
Adaptive surprisal	73.20	86.47

3.3 L1 Experiment 1: Modeling Human Expectations

Additionally, they tested whether the adaptive LM matches human expectations better than the non-adaptive model. To probe this, they tailored the adaptive LM to each story in the Natural Stories Corpus. After each story, they reverted to the initial Wikipedia-trained LM and restarted adaptation on the subsequent story., This model highly likely resulted a conservative estimate of the benefit of adaptation compared to the model that adapts continuously across multiple texts from the same story, similar to how humans might do.

They used surprisal to link a function between the NLM's predictions and human reading times. In general, surprisal means how unpredictable each word is given the preceding words, as represented below.

[1] Perplexity is the most widespread currency of evaluation for neural language models. Perplexity is defined as the inverse of the geometric average of the probability for each word (as defined in subsection 4.1). In general, a low perplexity indicates that the probability distribution is good at predicting upcoming words. Better language models will achieve lower perplexity scores or higher probability values on test sentences.

$$\text{surprisal}(w_i) = -\log \text{P}(w_i \mid w_1 ... w_{i-1})$$

They fitted the self-paced reading times in the Natural Stories Corpus with linear mixed effects models. In so doing, they reported that non-adaptive surprisal was a significant predictor of reading times ($p < 0.001$) when the model only included other baseline factors such as sentence position and word length, as shown in the top panel of Table 1. Adaptive surprisal was a significant predictor of reading times ($p < 0.001$) over non-adaptive surprisal and all baseline factors, as shown in the bottom panel of Table 2. Importantly, non-adaptive surprisal was no longer a significant predictor of reading times once adaptive surprisal was included. The results showed that the predictions of the adaptive model subsume the predictions of the non-adaptive one.

Table 2. Fixed Effect Regression Coefficients from Fitting Self-paced Reading Times
(taken from van Schijndel and Linzen 2018: 2)

	$\hat{\beta}$	$\hat{\sigma}$	t
WITHOUT ADAPTIVE SURPRISAL:			
Sentence position	0.55	0.53	1.03
Word length	7.29	1.00	7.26
Non-adaptive surprisal	6.64	0.68	9.79
WITH ADAPTIVE SURPRISAL:			
Sentence position	0.29	0.53	0.55
Word length	6.42	1.00	6.40
Non-adaptive surprisal	-0.89	0.68	-1.31
Adaptive surprisal	8.45	0.63	13.42

3.4 L1 Experiment 2: Reduced Relative Clauses

The foregoing experiment demonstrated that LM adaptation improves the ability to model the human expectations reflected in the self-paced reading time corpus. Furthermore, they further addressed the following two more questions: how much of improvement is due to the adaptation of the model's syntactic representations (Bacchiani et al. 2006, Dubey et al. 2006) and how much is simply due to the model assigning a higher probability to the words recently encountered (Kuhn and de Mori 1990, Church 2000)? They attempted to answer these questions using the two syntactic phenomena: reduced relative clauses and dative alternation in English. We first review the experiment concentrating on reduced relative clauses in English, as in (3).

(3) The grad students cheated during their exams left to go home.

The word 'cheated' in (3) is initially ambiguous between a main verb interpretation and a reduced relative clause interpretation. When the word 'left' is encountered, this ambiguity is resolved in favor of the reduced relative interpretation. Reduced relatives are an infrequent construction, and this condition makes the disambiguating word 'left' unexpected, So, readers read the sentence in (3) more slowly than they read the sentence in (4), where the words 'who were' signal early on that 'cheated' only is taken unambiguously as a passive verb inside the relative clause:

(4) The grad students who were cheated during their exams left to go home.

Fine and Jaeger (2016) reported that the cost of disambiguation in favor of the reduced relative interpretation decreased gradually with more exposures to reduced relative clauses. With more experiences of reduced relative clauses, readers can come to expect such clauses. Given this thesis, van Schijndel and Linzen (2018) adapted the model independently to random orderings of the critical and filler stimuli used in Experiment 3 of Fine and Jaeger (*ibid*.). Following them, van Schijndel and Linzen also used surprisal as a proxy for reading times and took the mean surprisal over three words in each ambiguous sentence: the disambiguating word and the following two words (e.g., ‘left to go’ in example (3) and (4)). To estimate the magnitude of the syntactic disambiguation penalty while also controlling for lexical content, they subtracted this quantity from the mean surprisal over the exact same words in the paired unambiguous sentence in (4).

To compare with the findings reported by Fine and Jaeger (2016), van Schijndel and Linzen (2018) replicated Fine and Jaeger’s method of plotting reading times as follows: they fitted a linear model of the mean surprisal of each disambiguating region with the number of trials that the model had been exposed to in the experiment at hand to account for a general trend of subjects/NLM speeding up over the course of the experiment. Then they plotted the mean residual model surprisal that was left in the diambiguating region in both the ambiguous and unambiguous conditions, as the experiment progressed. Fine and Jaeger’s results are illustrated in the top panel of Figure 1 and van Schijndel and Linzen’s results in the bottom panel of Figure 1.

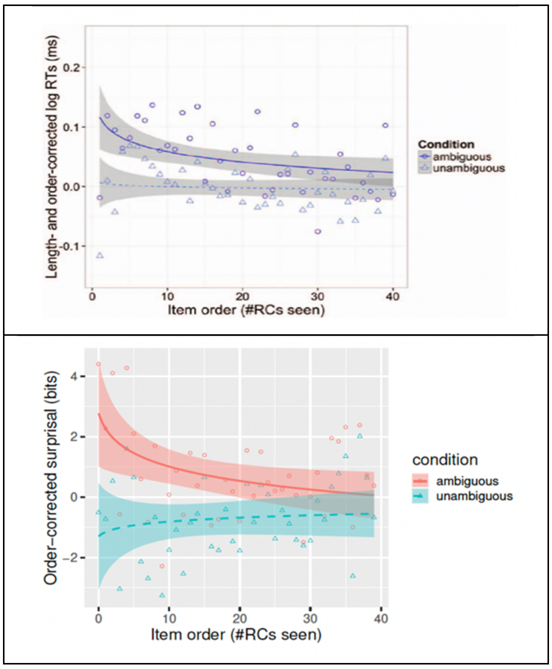

Figure 1. Mean Order-corrected Surprisal over the Disambiguating Region of the Critical Item
(taken from van Schijndel and Linzen 2018: 3)

The pattern of van Schijndel and Linzen’s results are identical to the human results reported by Find and Jaeger. Like humans, the model showed an initially large adaptation effect and a more gradual adaptation effect as the experiment progressed. Both humans and the NLM continued to adapt all the critical items more robustly over the course than just at the beginning phrase of the experiment. Besides, the model’s behaviors to unambiguous items did not change significantly throughout the experiment (p = .91).

3.5 L1 Experiment 3: Dative Alternations

In general, English has two roughly equivalent ways of expressing a transfer event:

(5) a. Prepositional Object (PO):
The man gave a ring to his girlfriend.
b. Double Object (DO):
The man gave his girlfriend a ring.

It is widely proposed that a recent exposure to one of these variants increases the probability of producing that variant (Bock 1986, Kaschak et al. 2006). To examine the LSTM LM's performance on the dative shift, van Schijdel and Linzen (2018) prepared 200 pairs of dative sentences, as exemplified in (5). They shuffled 100 DO sentences into 1000 filler sentences sampled from the Wikitext-2 training corpus (Merity et al. 2016). They then adapted the model to these 1100 sentences.

They froze the weights of the adapted model and tested its predictions for the two types of sentences. They also used the PO counterparts of the DO sentences in the adaptation set, which shared the vocabulary of the adaptation set but differed in syntax. Furthermore, they added 100 new DO sentences, which shared syntax but not content words with the adaptation set.

In addition, they explored the effect of learning rate in regard to LM adaptation. During adaptation, the model conducts a single parameter update after each sentence and does not train itself until it converges with a gradual reduction of learning rate, as would normally be the case during LM training. Consequently, the learning rate parameter decides the amount of adaptation that the model can undertake after each sentence[2]. This is why the optimal learning rate can vary from lexical adaptation to syntactic adaptation. Given this, the experiment manipulated learning rate on a logarithmic scale between 0.002 and 200. The results are illustrated in Figure 2.

As shown in Figure 2, the model successfully adapted to the DO construction as well as to the vocabulary of the adaptation sentences[3]. This behavior applied to all of the learning rates but for the learning rate of 200, which resulted in enormous perplexity in both sentence types. Both lexical and syntactic adaptation were greatly improved when the learning rate was around 2.

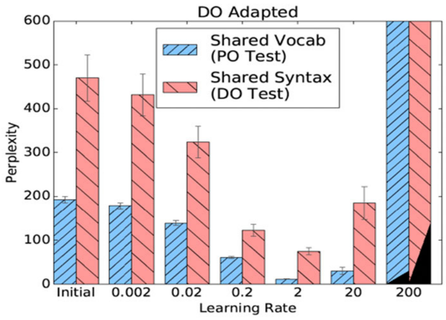

Figure 2. The Results of DO Adapted Model (taken from van Schijndel Linzen 2018: 4)

[2] If the learning rate is too low, adaptation will not have any effect; if it is too high, the model will overfit after each update and not generalize properly.

[3] For each bar graph, the left bar indicates the shared *vocab* sentences, while the right bar indicates the shared *syntax* sentences. This manipulation applies to Figure 3, Figure 5, and Figure 6.

Comparing learning rates of 2 and 20, syntactic adaptation was penalized at higher learning rates more severely than lexical adaptation. van Schijndel and Linzen analyzed these observed patterns in following ways: the DO adapting model can easily recognize the relevant vocabulary items, but syntax is not overtly recognizable. Syntactic properties must be inferred from multiple sentences of the similar syntax. As the learning made a process, a generalization was impeded by overfitting at higher learning rates.

Figure 3 reports the results of the PO adapting experiment. The patterns are similar to the results of the DO adapting experiment. The PO adapting model initially assigns a lower probability to the DO construction, compared with the syntactic adaptation test set (i.e., PO sentences). However, as the experiment made a progress, lexical adaptation was adequate to overcome this syntactic pre-training bias at the learning rate of 2 (e.g., the optimal learning rate).

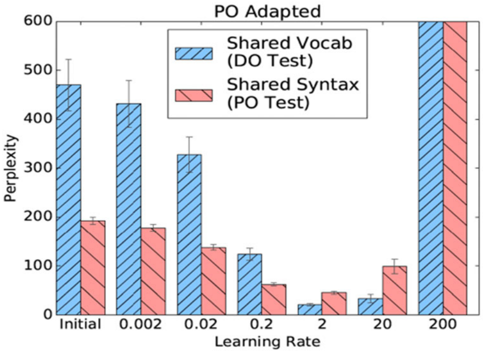

Figure 3. The Results of PO Adapted Model (taken from van Schijndel Linzen 2018: 8)

In short, van Schijndel and Linzen (2018) came up with a simple way of continuously adapting a neural LM based on the syntactic priming paradigm. Conducting experiments with controlled materials, they found that the LM adapts to both specific vocabulary items and abstract syntactic constructions.

Previous works including Schijdel and Linzen (2018) revealed that NLMs as well as humans can successfully adapt specific syntactic structures and assign a higher probability to recently learned structures. Based on the human and L1 NLM behaviors, we now start to examine how the L2 LSTM LM is sensitive to the syntactic priming paradigm, compared to its L1 counterpart. To investigate this, based on the pre-existing model, we implement the L2 LSTM model trained on the L2 corpus of English textbooks published in Korea. Applying the cumulative priming method, we aim to provide insight into the L2 adapted LM's representations of sentences containing reduced relative clauses or dative alternations in English. In the same fashion as Schijdel and Linzen (*ibid.*), we also adapt the L2 model to fairy tales and documentary genre from the Natural Stories Corpus (Futrell et al. 2018) to see how well the L2 adapted model expects upcoming words. In so doing, we explore what factors influence the strength of the priming effect for the L2 adapted LM. We now introduce how our experiments are designed to carry out the same tasks that van Schijndel and Linzen (2018) conducted.

4. L2 LSTM Language Model and Training Corpus

The model known as the Gulordava LSTM LM was built in the previous study to learn the English subject-verb number agreement task. This is the pre-existing LSTM LM trained on an L1 corpus of English sentences. In other words, it was pre-trained on 90 million tokens of English Wikipedia (Gulordava et al. 2018). Adopting the Gulordava model architecture, we implemented the L2 LSTM LM for our experiments. The implementation of the L2 LSTM language model was achieved on the training datasets collected from the EBS-CSAT English Prep Books published in 2016~2018 as well as from the English textbooks for Korean English L2ers based on the English 11 middle-school and L2 high-school textbooks published in Korea in 2001~2002 and on the English 19 middle-school and 12 high-school textbooks published in Korea in 2009~2010 (Choi and Park 2022a, Choi and Park 2022b, Choi et al. 2021, Kim 2019). In doing so, we gathered 7.9 million tokens from English textbooks for Korean L2 learners of English[4]. We also developed a data augmentation algorithm based on the textbook corpus to augment the training tokens to an additional dataset of 5.1 million tokens. This strategy allowed us to increase the amount of training data. In total, we collected a dataset of 13 million tokens for training the L2 LSTM LM. Using the L2 corpus, we trained the L2 LSTM, which is the baseline model. Following adaptation method mentioned in section 3.1, we further created three different L2 adapted LMs to conduct three different experiments. We will outline these adapted models in detail in the following subsections.

4.1 Perplexity for LSTM LM

As mentioned in Section 3, perplexity is the assessment metric to determine how good a language model is. Simply speaking, if the model has a perplexity of 50, it indicates that whenever the model is trying to predict the upcoming word, the model is as confused as if it has to pick between 50 words. In general, perplexity is defined as follows:

$$ppl(w) = 2^{-\frac{1}{n}\log_2 p(w_1, w_2, \ldots, w_n)},$$

The following Table 3 shows the perplexity of the two LSTM LMs at issue on the valid set. The L2 LSTM LM registers higher perplexity, compared to its L1 counterpart LSTM.

Table 3. The Perplexity of the Two LSTM LMs

LSTM LM	The Gulordava Model (L1)	L2
ppl	52.1	87.26

5. Results

5.1 L2 Experiment 1: Linguistic Accuracy

We begin by testing how well the L2 adapted model expects a next word. As its L1 counterpart NLM did, we report the L2 adapted model’s perplexity. We adapted the L2 LSTM LM to the first *x* sentence and then tested it on the *x* + 1 sentence for all *x*’s. We tested the L2 adapted model on the Natural Stories Corpus as mentioned in

[4] As mentioned above, we trained the L2 LSTM language model by employing 13 million tokens that Korean English learners may encounter in their English learning. The L2 LSTM LM was adopted for its previous success in learning a range of different syntactic structures such as filler-gap dependencies (Kim, 2019), linguistic anomalies (Choi et al., 2021), the dative alternation (Choi and Park, 2022a), relative clauses (Choi and Park, 2022b).

Section 3.1. We used the same test corpus to directly compare the adaptation effects[5].

As a result, the L2 adapted model did not improve test perplexity, compared to the L2 non-adapted model (986.59 vs. 979.19). Rather, the non-adaptive model registered slightly less perplexity than its adapted counterpart.

Next, we adapted the L2 LSTM LM to each genre separately. If the L2 adapted model tracks syntactic or stylistic patterns in one genre, we expect the adaptation effect to be more positive in the same type of genre. However, this prediction was achieved only in the documentary texts, but not in the fairy tales, as shown in Table 4. In other words, only the documentary texts displayed an improvement from adaptation. We interpret the results to show that text-specific adaptation was not always helpful even when the testing genre was similar to the training genre. Note that the L2 NLM was mainly trained on the textbooks for Korean L2 English learners. We assume that the aspects of the texts in the L2 corpus we used for training the L2 NLM may have affected the adaptation of fairy tales.

Table 4. The L2 LM's Test Perplexity

	Documentary texts	Fairy tales
Non-adaptive perplexity	1815.22	765.79
Adaptive perplexity	785.15	852.51

5.2 L2 Experiment 1: Modeling Human Expectations

Next, we tested whether the L2 adapted LM matches human expectations. We also tested the L2 NLM on the Natural Stories Corpus, which has 10 narratives with self-paced reading times from 181 native English speakers (Futrell et al., 2018). Given the corpus, we adapted the L2 LM to each story separately in the way that each reader saw the stories in a different order. After each story, we reverted to the initial Wikipedia-trained LM and restarted adaptation on the next story. In order to capture the relation between the L2 NLM's predictions and human reading times, we also used surprisal. We fitted the self-paced reading times in the Natural Stories Corpus with linear mixed effects models, a generalization of linear regression.

As with the L1 counterpart, non-adaptive surprisal with the L2 NLM was a significant predictor of reading times ($p < 0.0001$) when the model only included other baseline factors, as shown in the top panel of Table 5. Adaptive surprisal was a marginally significant predictor of reading times ($p = 0.053$) over non-adaptive surprisal and all baseline factors, as shown in the bottom panel of Table 5. The thing to note is that unlike the L1 counterpart, non-adaptive surprisal was a continuously significant predictor of reading times even when adaptive surprisal was included ($p < 0.0001$).

Table 5. Fixed Effect Regression Coefficients from Fitting Self-paced Reading Times

	Estimate	SE	t-value	p-value
Without Adaptive Surprisal				
Sentence position	0.39	0.48	0.80	0.212
Word length	10.65	0.96	11.08	< .0001
Non-adaptive surprisal	12.78	0.82	15.41	< .0001
With Adaptive Surprisal				
Sentence position	0.59	0.50	1.17	0.121
Word length	10.83	0.96	11.20	< .0001
Non-adaptive surprisal	13.07	0.84	15.40	< .0001
Adaptive surprisal	-0.63	0.39	-1.61	0.053

[5] In this study, we do not distinguish between priming and adaptation effects.

The results reported in Table 5 show that both the adaptive and the L2 non-adaptive LSTM LM exhibited significant priming effects on the text-specific experiment. Simply speaking, the predictions of the L2 adaptive model were similar to those of the non-adaptive model. Contrary to our expectation, adaptive surprisal was not superior to non-adaptive surprisal in modeling reading times.

5.3 L2 Experiment 2: Reduced Relative Clauses

The second experiment adapted the L2 model independently to random orderings of the critical (e.g., the reduced relative construction) and filler items employed in Experiment 3 of Fine and Jaeger (2016). One sample pair of the critical items are in Table 6 below. To replicate Fine and Jaeger's experiment, we gathered 40 critical items and 80 fillers into 16 lists (item orders). Four randomized orderings had the same items in each position as the first four but with opposite conditions for each critical item, and each of those eight total lists was presented in reverse order. Ambiguity was counterbalanced across experimental lists. Each stimulus list was presented in the exact same pseudorandom order. We used surprisal as a proxy for reading times and measured the mean surprisal over three words in each ambiguous sentence (e.g., *caught on right* as in Table 6).

Table 6. Examples of Ambiguous and Disambiguous Pair

Structure	Example
Ambiguous	The rookie technician taught the computer program caught on right away.
Disambiguous	The rookie technician who was taught the computer program caught on right away.

We directly compared the results from the L2 NLM with those reported by van Schijndel and Linzen. To see this, we followed their method of plotting reading times. As mentioned in the previous section, we fitted a linear model of the mean surprisal of each disambiguating region with the number of trials that the model had seen in the experiment at hand to account for a general trend of subjects/NLM speeding up over the course of the experiment. Then, we plotted the mean residual model surprisal that was left in the disambiguating region in both the ambiguous and unambiguous region conditions. We found that the pattern of the L2 adapted model results does not match those of both human and L1 adapted model results, as shown in Figure 4.

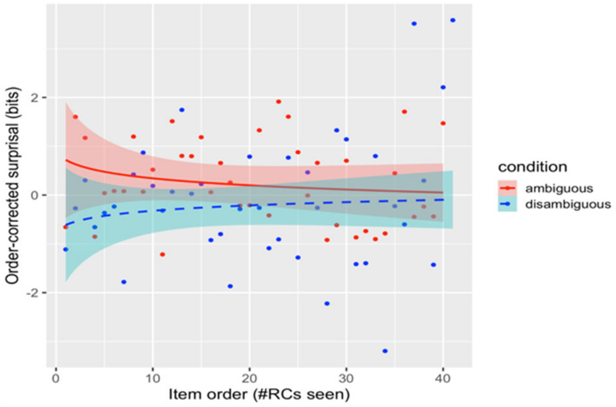

Figure 4. Mean Order-corrected Model Surprisal for the L2 Adaptive Model

The results from this experiment show that the L2 adapted model failed to yield positive adaptation effects. As the experiment progressed, the L2 model's responses to ambiguous items changed continuously but the adaptation effect was not statistically significant ($p = 0.43$). However, the mean surprisal values for the ambiguous items were marginally significant, relative to those for the disambiguous items ($p = 0.079$). Although the L2 adapted model exhibited a low adaptation effect, it distinguished the ambiguous and unambiguous conditions consistently throughout the experiment. In other words, the L2 NLM could encode the abstract structural information of reduced relative clauses quite reliably. We can infer that the sentences that belong to a linguistically interpretable class (e.g., sentences that match in ambiguity) were recognized similarly to each other in the L2 LM's representation space (Prasad et al., 2018). This is in line with the finding that the L2 adapted LM is capable of tracking abstract properties of sentences (Choi and Park 2022a, Choi and Park 2022b).

5.4 L2 Experiment 3: Dative Alternations

In the third experiment, we concentrated on the dative alternation. This phenomenon allows for the same content to be expressed by two different syntactic structures. The dative alternation includes ditransitive verbs whose complement can be expressed either by a double object (DO) structure or a prepositional object (PO) structure. We collected a number of DO and PO sentences, from which one sample pair is given in Table 7. All the experimental (i.e., critical and filler) items were taken from van Schijndel and Linzen (2018).

Table 7. Examples of Prime and Target Sentences

Structure	Example
Prepositional object (PO)	A man wrote a letter to a musician.
Double object (DO)	A man wrote a musician a letter.

As mentioned in section 3, for adaptation, 1,000 filler sentences were randomly drawn from the L2 training corpus, and 200 pairs of DO and PO sentences were generated. Then, the 200 pairs of DO and PO sentences were divided evenly into two sets. For each set, one of the DO and PO sentences was shuffled into the 1,000 filler sentences for adaptation, and the other was used to evaluate the L2 adapted model's ability to reproduce their variants.

In addition, we also examined the effect of learning rate on adaptation. Learning rate usually decides the amount of adaptation that the model can undertake after each sentence. If learning rate is too low, it is hard to find the adaptation effect. On the other hands, if learning rate is too high, the model may overfit after each update and fail to generalize. The optimal learning rate varies depending on the degree of lexical and syntactic adaptation.

Like van Schijndel and Linzen (2018), we predicted that the L2 NLM could adapt to DO/PO alternation as well as lexical items of the adaptation sentences. The prediction was fulfilled. As in Figure 5[6], the L2 LM adapted well both to DO constructions and the lexical items of the adaptation sentences. These behaviors were found in all of the learning rates but for 20 and 200, which resulted in billions of perplexity. Both syntactic and lexical adaptation were most successful when the learning rate was around 0.2.

[6] The results reported in Figure 5 and Figure 6 are taken from Choi and Park (2022a), who studied the same issue in this sub-section only. They employed the same L2 LSTM language model mentioned in Section 4 for their experiment.

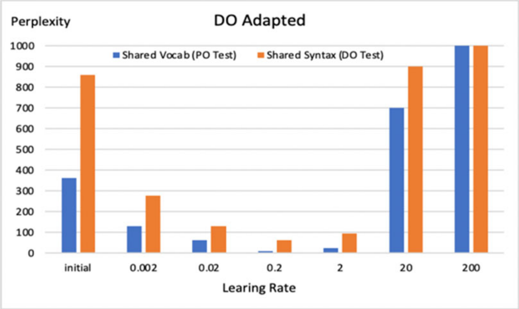

Figure 5. The Results of the L2 DO Adapting Model

Let us compare the learning rates from 0.2 to 20. Syntactic adaptation was penalized at higher learning rates more than lexical adaptation. Recall that this pattern was also discovered in the DO adapting L1 model, especially when comparing the learning rate of 2 with that of 20. As in van Schijndel and Linzen (2018), we also noted that tracking abstract lexical and syntactic properties from multiple similar sentences is impeded by overfitting at higher learning rates.

Now we report the results of the PO adaptation variants. In this case, the adaptation set contained PO sentences instead of DO sentences. Figure 6 illustrates that their results were similar to those of the DO adaptation set. Both syntactic and lexical adaptation were most successful when the learning rate was around 0.2, as found in the DO adapting model.

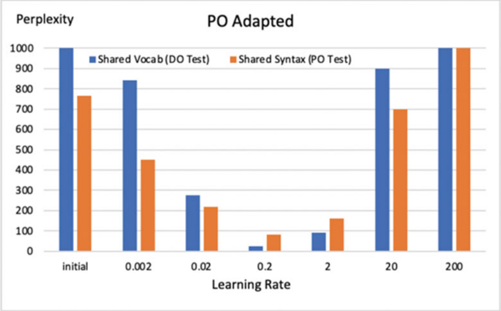

Figure 6. The Results of the L2 PO Adapting Model

At the beginning of this experiment, the PO adapting L2 model assigned a lower probability to the DO construction. This behavior was maintained when the learning rate was around 0.02. However, at the optimal learning rate (e.g., 0.2), lexical adaptation with DO sentences had to overcome this syntactic pre-training bias, just like the L1 counterpart. Interestingly, what is also worth mentioning is that the learning rate of 20 resulted in billions of perplexity again, especially with the DO test set. This peculiar pattern was also observed in the DO adapting L2 model, as illustrated in Figure 5.

Taken together, the results of the DO & PO adapting L2 LM are similar to those of the DO & PO adapting L1 LM, reported in van Schijndel and Linzen (2018). The L2 LM appropriately adapted lexical items as well as abstract syntactic structures. The DO & PO adapting L2 LM exhibited positive priming effects with the English dative alternation construction, compared with reduced relative clauses. Furthermore, we found that learning rate crucially affects both lexical and syntactic adaptation. L1 and L2 adapted language models were similar regarding the effect of learning rate on adaption. However, the optimal learning rate was different: the DO & PO adapting L1 model is around 2, while the DO & PO adapting L2model is around 0.2. This means that the L2 adapted LM assigned greater probabilities to the items that in fact occurred in prior contexts, than the L1 adapted language model did.

6. Discussion

In this study, we tried to answer the following two research questions: (1) Are the L2 adapted language model susceptible to syntactic priming like its L1 counterpart? (2) Which environmental factors influence the L2 priming effect acquired by the L2 adaptive model? To examine these two issues, we replicated the previous experiments in van Schijndel and Linzen (2018), running a series of computational experiments on several different types of syntactic adaptation, and compared the results from the L2 and the L1 adapted language models.

First, the current study did not find a consistent improvement on word prediction accuracy with the L2 adapted LM. We reported that only the documentary texts exhibited an improvement from adaptation. However, text specific adaptation was not helpful with the fairy tales. It did not positively impact on priming in the case of the fairy tales, even when the testing genre was similar to the training genre. We then examined whether the L2 adaptive LM correlated with human expectations more significantly than its non-adaptive counterpart. Both the L2 adaptive and the L2 non-adaptive models turned out to be a significant predictor of reading times. The result indicates that the predictions of the adaptive model were not differentiated from those of the non-adaptive model. These two L2 LMs played a role in determining its probability in line with human expectations.

In the second experiment, we adapted the L2 NLM independently to random orderings of the sentences containing reduced relative clauses. Using surprisal, we measured the mean value over the three words in each ambiguous sentence. We directly compared the adaptation effects in the L2 NLM and its L1 LM counterpart. The L2 adaptive model did not record a significant adaptation effect. As the experiment continued, the L2 model's behavior to ambiguous items changed constantly, but the adaptation effect was not statistically significant. Still, the average surprisal values of the ambiguous items were marginally significant than those of the disambiguous items. The observed behavior of the L2 NLM indicates that it has an ability to track abstract linguistic properties quite reliably. We have compelling evidence that the model encodes information about the ambiguity features of reduced relative clauses.

In the third experiment, the results for the DO & PO adapting L2 LM were similar to those of the DO & PO adapting L1 LM in van Schijndel and Linzen (2018). The L2 adapted LM registered priming effects with the English dative alternation construction. Furthermore, we noted that learning rate influenced both lexical and syntactic adaptation. L1 and L2 adapted LMs were similar concerning the general effect of learning rate on adaption, but they differed in the optimal learning rate.

Overall, our study has revealed novel details about the nature of the representations learned by the L2 NLM about syntactic aspects of sentences. The findings from the study of the L2 NLM have demonstrated the usefulness of the priming paradigm for investigating the research questions raised above. We have particularly noted that a quite high degree of abstract syntactic structure is being represented by the L2 NLM. Even though we have failed to discover adaptation effects among all of the three experiments, we have shown the benefits of repurposing the priming paradigm to compare the degrees of the knowledge acquired by the two different L1 and L2 NLMs. It is to be underscored that the amount of training data for an NLM may have given rise to quantitatively different effects between L1 and L2 NLMs. In addition, as for humans, priming effects for an NLM can be affected by various aspects of data other than its amount. In this sense, different environmental factors may influence an NLM's performance in priming.

7. Conclusion

In this research, we have probed the syntactic priming behaviors of the L2 LSTM LM. We have evaluated how well the L2 LSTM LM expects a greater probability of syntactic structures that it has potentially learned. We have provided further evidence that the L2 adaptive LM is able to track abstract linguistic properties of critical items in line with previous works. In addition, based on the materials that teased apart lexical content from syntax, we have demonstrated that the L2 LSTM LM can adapt both lexical and syntactic predictions. Even though the L2 adaptive model exhibited low adaptation effects in the text-specific experiment and reduced relative constructions, we concurred with the claim that the cumulative priming method is a helpful and effective way of investigating the nature of the internal representations of specific syntactic structures. Following the priming method, we have examined the correlation between surprisal values derived from the L2 NLM and human reading times.

Before concluding, we mention some preliminary remarks on the importance of training dataset size in investigating the nature of NLMs. Since NLMs are data-driven, we need to investigate deeply how the amount of training data influences the performance of NLMs like the LSTM LM. By comparing them we can contribute to a better understanding of the differences between them, like those between the L1 and the L2 LSTM LMs reported in this paper. Though we have reported a preliminary study of them here, a more thorough examination of them is left for future study.

References

Bacchiani, M., M. Riley, B. Roark and R. Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech & Language* 20(1), 41-68.

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

Bock, J. K. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3), 355-387.

Bock, J. K. and Z. M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General* 129(2), 177.

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

Choi, S. J. and M. K. Park. 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 M. K. Park. 2022b. Syntactic priming by L2 LSTM language models. *The Journal of Studies in Language* 37(4). 475-489.

Church, K. 2000. Empirical estimates of adaptation: The chance of two noriegas is closer to p/2 than p2. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Davis, F. and M. Van Schijndel. 2020. Recurrent neural network language models always learn English-like relative clause attachment. *arXiv preprint arXiv:2005.00165*.

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

Fine, A. B. and T. F. Jaeger. 2016. The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42(9), 1362.

Futrell, R., E. Gibson, H. Tily, I. Blank, A. Vishnevetsky, S. T. Piantadosi and E. Fedorenko. 2017. The natural stories corpus. *arXiv preprint arXiv:1708.05763*.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Kaschak, M. P., R. A. Loney and K. L. Borreggine. 2006. Recent experience affects the strength of structural priming. *Cognition* 99(3), B73-B82.

Kaschak, M. P., T. J. Kutta and C. Schatschneider. 2011. Long-term cumulative structural priming persists for (at least) one week. *Memory & Cognition 39*(3), 381-388.

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.

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(6), 570-583.

Mahowald, K., A. James, R. Futrell and E. Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91, 5-27.

Merity, S., C. Xiong, J. Bradbury and R. Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Pickering, M. J. and H. P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language* 39(4), 633-651.

Pickering, M. J. and V. S. Ferreira. 2008. Structural priming: a critical review. *Psychological Bulletin* 134(3), 427.

Prasad, G., M. Van Schijndel and T. Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. *arXiv preprint arXiv:1909.10579*.

Ravfogel, S., G. Prasad, T. Linzen and Y. Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. *arXiv preprint arXiv:2105.06965.*

Sinclair, A., J. Jumelet, W. Zuidema and R. Fernández. 2021. Syntactic persistence in language models: Priming as a window into abstract language representations. *arXiv preprint arXiv:2109.14989*.

Tenney, I., D. Das and E. Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

van Schijndel, M. and T. Linzen. 2018. A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930*.

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

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