Sign Languages and Linguistics Convergence Research
Vol. 1. No. 1 (2024) pp.058~086

Grammatical Transfer in L2 Neural Language Models

Park, Myung-Kwan
(Dongguk University)

Park, Myung-Kwan. 2024. Grammatical Transfer in L2 Neural Language Models. ***Sign Languages and Linguistics Convergence Research,*** *1*(1), 58-86. This study examines the impacts of initial grammatical knowledge from the first language (English) on the acquisition of grammatical structures in the second language (Korean) by neural language models (NLMs). To this end, we employ the standard experimental procedure in the rapidly-evolving field of natural language processing (NLP) that encompasses (i) pre-training Transformer-based GPT-2 NLMs on either Korean or English datasets, (ii) further fine-tuning them using a targeted set of Korean data as either the first or the second language, and (iii) assessing their performances using the K-BLiMP test data while analyzing their grammatical generalization capabilities. The findings in this study indicate that negative transfer effects arise in L2 NLMs when comparing models that learn Korean as the second language with those that learn Korean as the first language. Additionally, during the progression of training, the NLMs learning the second language attained the acquisition of its grammatical features, but to different degrees depending on the construction types tested. This study enhances our understanding of potential grammatical difficulties that may occur due to the interference from the first language when NLMs learn Korean as the second language.

Key words: Neural Language Model, L2 Learning, Grammatical Transfer, L1 Interference, Grammatical Features

1. Introduction

Cross-linguistic transfer is a phenomenon observed when the linguistic structures and knowledge of a person's first/native language (L1) exert influence on their ability to learn and use a second, foreign language (L2). This influence can manifest in multiple facets of language acquisition, including vocabulary, grammar, pronunciation, and syntax.

The underlying mechanism of cross-linguistic transfer involves the cognitive processes by which a language learner applies familiar linguistic patterns from their first language to a new language. Essentially, when faced with the task of learning and using a second language, the brain often relies on existing linguistic knowledge as a point of reference. This can lead to transfer effects, which are changes in the way that the second language is learned or expressed due to the influence of the first language.

Transfer effects can be classified into two main types: positive transfer and negative transfer. Positive transfer occurs when knowledge of the first language aids in learning a similar aspect in the second language. For example, a native Spanish speaker may find it easier to learn Italian because of the similarities in vocabulary and grammar between the two Romance languages. This facilitation can speed up language acquisition and enhance fluency.

On the other hand, negative transfer, often referred to as interference, arises when the differences between the first and the second languages hinder learning and lead to errors. This is commonly seen in the misuse of grammatical structures or pronunciation. For instance, a French speaker learning English might struggle with the placement of adjectives because, in French, adjectives typically follow the noun, whereas in English, they usually precede it.

In the realm of second language acquisition and learning (SLA/L), significant research has been dedicated to exploring patterns of negative transfer. These patterns are crucial as they not only illuminate the complexities of learning a second language but also provide insights into the typological relationship between a first language and a second language. For instance, Dulay and Burt's (1974) study on children learning a second language identified specific errors linked to negative transfer from their first language, alongside the errors

stemming from cognitive strategies similar to those used in first language acquisition, termed developmental errors. Similarly, Berzak et al. (2014) discovered strong correlations between the structural similarities in English as a Second Language (ESL) texts and the typological characteristics of the learners' first languages, further highlighting structural similarities in grammar between languages like Korean and Japanese.

Recent studies have also focused on grammatical transfer in neural language models (NLMs), with significant contributions from research on Transformer-based NLMs featuring multi-head attention mechanisms and transfer learning via pre-trained models (Vaswani et al., 2017; Ruder et al., 2019; Conneau et al., 2017, 2018; Artetxe et al., 2018; Wu and Dredze, 2019; Wu et al., 2019). These models have become instrumental in advancing state-of-the-art performances in various Natural Language Processing (NLP) tasks, with self-supervised learning objectives now being standard in the field (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2018; Ruder, 2019).

Likewise, from a scientific and engineering viewpoint, there has been extensive exploration into the mechanisms and linguistic properties of language transfer in NLMs (Papadimitriou and Jurafsky, 2020; Perez-Mayos et al., 2021; Blevins et al., 2022; Ri and Tsuruoka, 2022; Tyler et al., 2022). This includes the studies extending beyond natural language transfer to applications inspired by pedagogical objectives, such as simulating second language learning (Settles et al., 2018). Notably, large English NLMs have demonstrated translation capabilities with minimal training data in non-English languages, indicating successful language transfer from English to other languages (Brown et al., 2020; Shi et al., 2022). The efficacy of these cross-lingual transfers has been assessed using detailed metrics like perplexity and task-specific accuracy (Papadimitriou and Jurafsky, 2020; Blevins et al., 2022; Deshpande et al., 2021).

Furthermore, recent research has increasingly focused on the role of pre-trained NLMs in learning grammatical structures and understanding language transfer dynamics in multilingual contexts. Studies by Yadavalli et al. (2023) and Oba et al. (2023) have thoroughly investigated these aspects, identifying both positive and negative transfer effects and how an individual’s first language structure impacts their ability to learn a second language.

Based on these previous studies, this study also to further explore these grammatical transfer phenomena within NLMs in the context of second language learning. Particularly in this paper, our primary research focus is to examine how grammatical knowledge from the first language (English) influences the learning of grammatical structures in the second language (Korean) within NLMs. To explore this issue, we adopt a well-established experimental approach detailed in Section 3. This includes (i) pre-training NLMs on either Korean or English datasets, (ii) fine-tuning these models specifically on Korean as either the first (L1) or the second language (L2), and (iii) evaluating and analyzing their capacities for grammatical generalization across the two language contexts. Due to the constraints on data size and availability as an open source, we employ a Transformer-based model, GPT-2, rather than larger models like GPT-3/4.

The structure of the paper is organized as follows: Section 2 reviews previous research on cross-linguistic transfer in NLMs. Section 3 describes our methodology and the development of NLMs, including specifics on the training and testing data, as well as the experimental procedures employed. Section 4 presents the results of our experiments concerning grammatical knowledge in the NLM tested in this study. Section 5 discusses the implications of our findings on grammatical transfer. Section 6 concludes the paper with a summary of our insights and final thoughts.

2. Previous Studies

The exploration of language learning capabilities within NLMs has garnered considerable attention, notably for the strides made by these NLMs in various linguistic tasks. Recently, a particular focus has been funneled on how these NLMs acquire and process grammatical knowledge, especially through the lens of second language acquisition. This shift towards understanding the dynamics of language transfer — how knowledge of a first language influences learning a second language — marks a significant evolution in the study of pre-trained NLMs.

Studies like those conducted by Yadavalli et al. (2023) and Oba et al.

(2023) exemplify this new direction. Oba et al. (2023) expanded the scope of research to bilingual NLMs, specifically investigating the acquisition of a second language. Their research illuminated the accelerated learning in L2, facilitated by prior training in various first languages such as French, German, Japanese, and Russian. This pre-training in L1 not only quickened the linguistic generalization in English as L2 but also revealed that the nature of linguistic transfer can profoundly impact these generalizations. Interestingly, their experiments suggested that L2 grammar acquisition was somewhat hindered when NLMs were fed translation pairs from L1 to L2, as opposed to exclusive exposure to L2 monolingual texts. Their findings also indicated that the experiences with L1 significantly enhance linguistic generalization capabilities in L2. At the same time, the specific characteristics of L1 can determine the generalization patterns in L2, echoing the observations made by Chiswick and Miller (2005) about the varying difficulty levels of linguistic transfer.

Yadavalli et al. (2023) addressed a gap in the study of linguistic transfer in NLP tasks by examining both positive and negative transfer effects between first and second languages. Their study involved sequential second language learning in NLMs, utilizing a dataset comprising five typologically diverse languages — German, French, Polish, Indonesian, and Japanese — with English as the second language. The NLMs were pre-trained on native Child-Directed Speech (CDS), closely mirroring the linguistic input received by young children. This approach, grounded in cross-lingual transfer learning techniques proposed by Papadimitriou and Jurafsky (2020), emphasized conversational speech data's superiority over scripted speech in fostering effective language acquisition. Utilizing the BLiMP (Benchmark of Linguistic Minimal Pairs) grammar test suite (Warstadt et al., 2020), their findings revealed a negative correlation between the typological distance of the CDS language from English and the performance of the L2 NLMs. This suggests that greater linguistic differences between an NLM's L1 and L2 can impede its ability to perform well in L2.

In our current study, we build on these insights by pre-training NLMs using English texts as L1 and Korean as L2, aiming to further identity the nuances of grammatical transfer. We delve deeper into how specific grammatical features transferred from English influence the acquisition and processing of Korean,

providing a richer understanding of cross-linguistic influence in NLMs. We now turn to the next section to report and discuss the detailed outcomes of our experiment.

3. Acquisition of L2 Korean in NLMs

To explore how knowledge of a first language (L1) influences grammatical generalization in a second language (L2) in NLMs, we implemented an experimental procedure schematized in Figure 1. Initially, we simulated L1 acquisition by pre-training NLMs on separate monolingual corpora for English and Korean. Following this, we simulated L2 acquisition by further fine-tuning these pre-trained NLMs on a corpus consisting of Korean texts that were compiled from contemporary short stories/novels and essays on the Internet. The final stage of our experiment involved assessing the NLMs' abilities to make grammatical judgments in Korean, using the K-BLiMP (Korean Benchmark of Linguistic Minimal Pairs) test suite, which will be further described in Section 3.2.

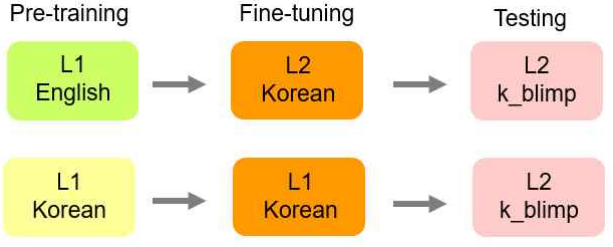


Figure 1: Experimental Procedure

In this study, we employed a Transformer-based NLM, GPT-2 (Generative

Pre-trained Transformer 2; OpenAI).[1] GPT-2, a large-scale NLM built on the Transformer architecture (Vaswani et al., 2017), has been shown to perform exceptionally well on the (E-)BLiMP test suite, outperforming other NLMs according to Warstadt et al. (2020). This NLM processes input sequentially from left to right, mirroring human language processing. Our selection of GPT-2 for this study is due to its proficiency in analyzing grammatical structures, making it an ideal candidate for investigating the cross-linguistic transfer effects from English to Korean in NLMs.[2]

In this study focused on first and second language acquisition, pre-trained NLMs from Hugging Face's extensive repository were pivotal. For the Korean (L1) component, the KoGPT-2 model, developed by SK Telecom, was chosen for its specialized training on comprehensive 40GB datasets (that amounts to about 6.67 billion tokens) composed of diverse Korean texts. These datasets were meticulously assembled to represent a broad spectrum of modern Korean usage, encompassing genres such as news articles, literary works, and varied web texts. This variety ensures that this NLM has a broad linguistic base, reflecting the complexities and nuances of contemporary Korean language.

Similarly, for the English (L1) acquisition, the GPT-2 model developed by OpenAI was employed. GPT-2's training was conducted on the web-text dataset, an extensive 45GB collection (about 7.5 billion tokens) that OpenAI curated by scraping a wide range of content from the Internet. This dataset includes a heterogeneous mix of articles, essays, and other written materials that provide a rich linguistic environment, mirroring the diverse and multifaceted nature of English usage across different contexts and subjects. Such a dataset ensures that GPT-2 has exposure to a vast array of lexical choices and syntactic structures typical of native English speakers.

---

1) We developed our NLMs based on ttps://github.com/huggingface/transformers/ tree/mai n/examples/pytorch/language-modeling.

2) Although GPT models are primarily designed as generative models, their generative capabilities are closely linked to their understanding of language. Therefore, the findings of this study could positively apply to the perspective of language understanding as well. Effective text generation requires deep language comprehension, suggesting that the impact observed in generative models is likely to extend to language understanding models as well.

Moving to the aspect of second language acquisition, the research leveraged the English GPT-2 model, which was fine-tuned using a corpus of carefully selected Korean short stories/novels and essays. This corpus, totaling approximately one million words, is crucial for its role in equipping the NLM to process and understand cross-linguistic information. The inclusion of short stories/novels and essays not only introduces complex narrative structures to the L2 NLM but also immerses it in culturally rich contexts that challenge and enhance its language understanding capabilities.

This fine-tuning is integral to adapting the English-trained, and then Korean-finetuned GPT-2 model to effectively function in a Korean linguistic setting, thereby facilitating a deeper investigation into bilingual language acquisition dynamics. By analyzing how the L2 NLM handles the transition from English to Korean, insights can be gleaned about the cognitive and linguistic processes involved in learning a second language, offering valuable implications for both linguistic theory and AI development. This strategic use of sophisticated L2 NLMs exemplifies how artificial intelligence can simulate and study complex human language learning behaviors in a controlled, replicable manner.

The L2 NLM in focus in our study underwent a comprehensive fine-tuning process consisting of 25,000 steps. This extensive fine-tuning phase is crucial as it helps the L2 NLM gradually adapt from its initial pre-trained state (on the first language, L1 data) to effectively handle the L2 data, in this case, Korean. The fine-tuning is designed to mimic the way that humans might learn a second language, progressively building upon existing knowledge and adjusting to new linguistic patterns and rules.

To analyze this progression in greater details, we established intermediate checkpoints at specified intervals during the fine-tuning process. At each of these checkpoints, the L2 NLM's current state was evaluated using the K-BLiMP test suite. K-BLiMP is specifically designed to assess the grammatical judgment capabilities of L2 Korean NLMs by presenting them with pairs of sentences that differ in minimal ways, testing various aspects of Korean grammar. By examining how the L2 NLM performs on these tests at different stages of training, we can gain insights into how its understanding of Korean grammar evolves and improves over time.

This methodological approach not only allows us to capture snapshots of the L2 NLM's learning at various points but also helps in identifying specific areas and stages where the L2 NLM may struggle or excel in acquiring L2. It provides a nuanced understanding of the learning process, illustrating the gradual and sometimes nonlinear nature of language acquisition in an L2 NLM. This, in turn, informs our broader research objectives regarding cross-linguistic transfer effects and the capabilities of an L2 NLM in learning grammatical structures across different languages.

3.2 Evaluation Set

The evaluation of the L2 Korean NLM's understanding of the Korean language was conducted using the K-BLiMP, a specially designed dataset aimed at conducting syntactic assessments of the NLM concerned. This benchmark was inspired by similar frameworks in other languages: (E-)BLiMP for English and J-BLiMP for Japanese, both of which are tools used to assess grammatical judgment capabilities in their respective languages. K-BLiMP is comprised of 3,350 minimal pairs; these are pairs of sentences that are almost identical but differ slightly in a way that affects grammaticality, allowing for precise testing of language/grammar understanding.

The K-BLiMP dataset is organized into 14 distinct groups, each dedicated to a particular grammatical phenomenon and categorized under one of three broader grammatical domains: (i) morpho-syntax (Case Particles; Tense/Aspect/Modal; Honorifics; NP Structure); (ii) syntax (Relatives; Question Formation and Islands; Displacement; Clefts; Adverbial and Coordinate Clause Formation; Actives vs. Passives); (iii) the syntax-semantics interface (NPIs/NCIs; Negations; Argument Structure; Adjectives vs. Adverbs). This structured arrangement facilitates a focused analysis of the L2 Korean NLM's competencies across different aspects of English and Korean, making it an effective tool for identifying both strengths and weaknesses in the L2 Korean NLM's grasp of Korean.

To quantitatively measure the influence of the first language's structure on the acquisition of the second language in the NLM concerned, we employed

sentence perplexity as a metric, following the methodology outlined by Warstadt et al. (2020). Sentence perplexity evaluates how well the NLM predicts a given text, with lower perplexity indicating that an NLM finds the text more predictable, which can be interpreted as a sign of better understanding. This metric is particularly useful for assessing the NLM's fluency and grasp of grammatical norms in the target language.

In addition to perplexity, we also computed the macro-average accuracies across all the test sets included in K-BLiMP. This involves calculating the average accuracy for each of the 14 groups separately, and then averaging these figures to get the macro-average. This metric provides a broad view of the L2 Korean NLM's overall performance and its ability to generalize across different grammatical phenomena. Together, these analyses offer a comprehensive picture of how effectively the L2 Korean NLM concerned can transfer knowledge from its first language and apply the knowledge to a new, second language, highlighting the specific areas where transfer is successful and where it may fall short.

4. Experimental Results

In the experiment in this paper, the accuracy of the fine-tuned NLMs was determined by measuring the percentage of cases in which the models assigned a higher probability to the grammatically correct sentence over the grammatically incorrect one, using the K-BLiMP minimal pairs. The performance comparison of the two NLMs was based on accuracy scores obtained from 14 distinct grammatical tests provided by the K-BLiMP test suite, as detailed in section 4.1 of this paper. Additionally, the learning trajectory of the NLMs regarding second language (L2) acquisition was outlined in section 4.2.

4.1 Overall Accuracy

This section provides a detailed evaluation of the NLMs in the investigation of the issues in this paper, showcasing the accuracy scores through tables and

figures, presented as percentages for clear understanding. The aggregated average accuracy results for all the NLMs are displayed in Table 1 and Figure 2.

Table 1 highlights the fact that the pre-trained KoGPT-2 achieved the highest overall accuracy on the K-BLiMP benchmark, benefiting from its extensive training on a large Korean dataset. In contrast, the En(L1)-Ko(Fine-Tuned(FT)-L2) GPT-2 model, which was initially trained in English and then fine-tuned with Korean, showed the lowest accuracy. This drop is attributed to the challenges associated with mastering L2 grammatical structures when Korean is learned as the second language after English as the first language.

The accuracy of the Only-Ko(FT-L1) GPT-2 model surpassed that of the En(L1)-Ko(FT-L2) GPT-2 model, suggesting a little better performance when the former was not previously exposed to English datasets. The latter lagged a bit behind the former, which was directly fine-tuned in Korean as the first language, and largely behind the pre-trained KoGPT-2, which was exposed to substantial Korean datasets from the outset.

Statistical analysis, including the t-test, indicates no significant difference in the performance between the En(L1)-Ko(FT-L2) and Only-Ko(FT-L1) models ($p=0.41$). However a significant difference was noted when comparing En(L1)-Ko(FT-L2) with Ko(L1) ($p<0.05$). This finding suggests that the interplay between L1 and L2 language training has significantly impacted model performance, warranting further research into the dynamics of grammatical transfer.[3)]

Table 1. Average Accuracy of Each Model on the K-BLiMP Test Suite[4)]

3) We report no significant difference between the En(L1)-Ko(FT-L2) and Only-Ko(FT-L1) models, but did find a significant difference between the En(L1)-Ko(FT-L2) and Ko(L1) models, suggesting the interplay between L1 and L2 language training impacts model performance. However, one anonymous reviewer notes that the lack of better performance in the En(L1)-Ko(FT-L2) model compared to the Only-Ko(FT-L1) model raises questions. It suggests that the significant difference observed might be due to the presence of Korean in the fine-tuning data rather than negative transfer effects from L1 and L2 interplay. See Section 4.2 for the learning trajectories of the models, which addresses this issue.

	Ko(L1)	Ko(L1)-Ko(FT-L1)	En(L1)-Ko(FT-L2)	Only-Ko(FT-L1)
Average	81.78	74.27	58.33	59.11

To assess the grammatical proficiency of each model in the 14 linguistic phenomena, we calculated the accuracy for each phenomenon using on the K-BLiMP test suite, as shown in Figure 2 (where the Y-axis represents the average accuracy scores). An analysis of these grammatical phenomena reveals that the performance of En(L1)-Ko(FT-L2) was below chance level in 5 out of the 14 evaluated phenomena in the K-BLiMP test suite. Conversely, the NLMs trained with Korean as L1 demonstrated superior performance across nearly all the tested phenomena.

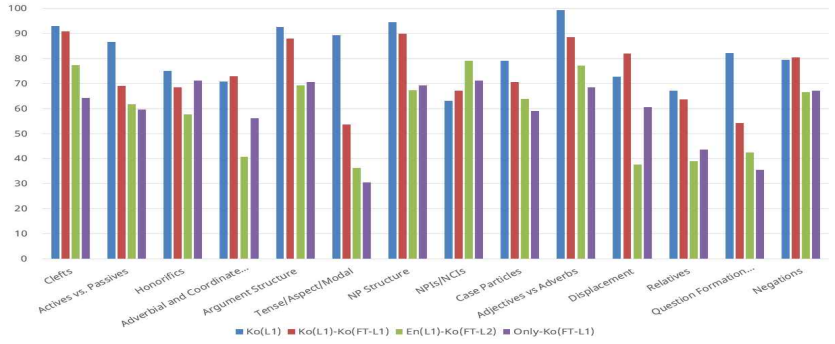


Figure 2. Each Model's Average Accuracy for Each Phenomenon on the K-BLiMP

4) On top of the NLMs reported in the text, the average score of SKT-KoBERT is 56.73, and that of KLUE-RoBERTa is 81.72. Each score of the 14 grammatical phenomena for these two models is shown in the Appendix.

4.1.1 Comparison between Ko(L1) and En(L1)-Ko(FT-L2)

To examine the cross-linguistic transfer effects between English as the first language and Korean as the second language, we compared the En(L1)-Ko(FT-L2) model with the Ko(L1) model, focusing on the accuracy of each grammatical phenomenon. The En(L1)-Ko(FT-L2) model, which learned Korean as the second language, demonstrated lower performance across all the phenomena compared to the Ko(L1) model. Figure 3 illustrates the contrast between En(L1)-Ko(FT-L2) and Ko(L1), revealing substantial drops in performance on the part of the former, compared to the latter.

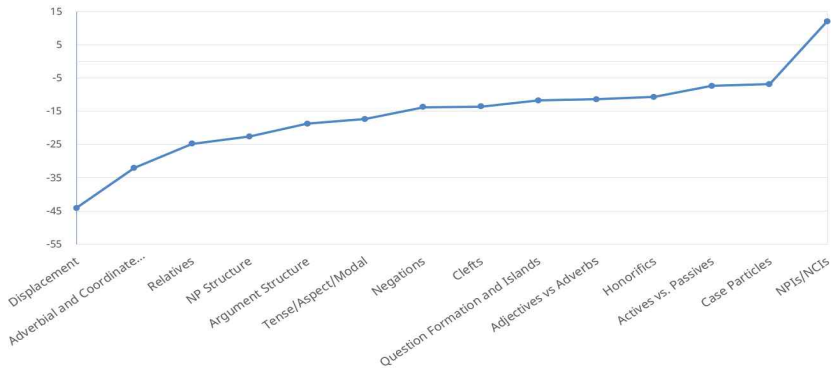


Figure 3. Performance Drop in the En(L1)-Ko(FT-L2) Model, compared to the Ko(L1) Counterpart, for the 14 Phenomena

Specifically, a performance drop of more than 10 points was observed in the 11 phenomena except 'Actives vs. Passives', 'Case Particles', and 'NPIs/NCIs'. The most significant decline was in 'Displacement', with a decrease of 44.2 points. Additionally, notable decreases were seen in 'Adverbial and Coordinate Clause Formation' (32.1 points), 'Relatives' (24.8 points), 'NP Structure' (22.6 points), 'Argument Structure' (18.7 points), and 'Tense/Aspect/Modal' (17.3 points). These negative values for the En(L1)-Ko(FT-L2) model indicate negative transfer effects. An unusual behavior was noted in 'NPIs/NCIs', where performance increased by 12.1 points in the En(L1)-Ko(FT-L2) model.

To confirm grammatical transfer effects, we conducted a t-test comparing the

accuracy of each phenomenon in En(L1)-Ko(FT-L2) and Ko(L1)-Ko(FT-L1). The results of the paired t-test are summarized in Table 2. Notably, the differences between En(L1)-Ko(FT-L2) and Ko(L1)-Ko(FT-L1) for each phenomenon indicate a significant negative effect, with the exception of 'NPIs/NCIs' (all p's < 0.001).[5)]

Table 2. The Results of the paired t-test between the En(L1)-Ko(FT-L2) and Ko(L1)-Ko(FT-L1) Models

Phenomena	t−value	p−value
Case Particles	−10.77	***
Realtives	−15.65	***
Argument Structure	−6.41	***
Actives vs. Passives	−22.59	***
NP Structure	−63.34	***
Adverbial and Coordinate Clause Formation	−53.95	***
NPIs/NCIs	29.52	***
Negations	−13.53	***
Clefts	−8.25	***
Tense/Aspect/Modal	−16.78	***
Question Formation and Islands	−10.71	***
Displacement	−25.32	***
Honorifics	−12.06	***
Adjectives vs. Adverbs	−9.68	***

'***' p< 0.001 '**' p< 0.01 '*' p< 0.05

4.1.2 Transfer Effects in Three Categories

To investigate the cross-lingual transfer effects of our NLMs, particularly focusing on grammatical transfer effects, we categorized the 14 phenomena into three coarse groups: morpho-syntax, syntax, and the syntax-semantics interface, as shown in Table 3.

5) For the t-test and other statistical analyses below, we use the Ko(L1)-Ko(FT-L1) NLM rather than the Ko(L1) NLM, since the former involves fine-tuning like the En(L1)-Ko(FT-L2) NLM.

Table 3. 11 Phenomena Classified into Three Categories

Morpho-Syntax	Case Particles
	Tense/Aspect/Modal
	Honorifics
	NP Structure
Syntax	Relatives
	Question Formation and Island
	Displacement
	Clefts
	Adverbial and Coordinate Clause Formation
	Actives vs. Passives
Syntax-Semantics	NPIs/NCIs
	Negations
	Argument Structure
	Adjectives vs Adverbs

Although verbal agreement in 'Honorifics' and nominal structure in 'NP Structure' are considered relatively straightforward morpho-syntactic phenomena, Figure 4 shows a significant under-performance of En(L1)-Ko(FT-L2) compared to Ko(L1) on these tasks other than those involving 'Case Particles' and 'Tense/Aspect/Modal'. This finding suggests that the initial exposure to English morpho-syntactic rules, learned through English Wikipedia, might hinder the En(L1)-Ko(FT-L2) NLM from adapting to similar rules in Korean. This difficulty could be due to the increased complexity or variability present in Korean morpho-syntax.

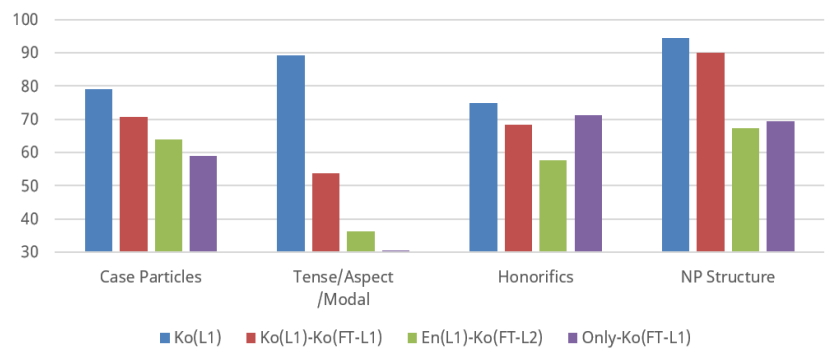


Figure 4. Each Model's Average Accuracy for Each Morpho-Syntactic Phenomenon

Figure 5 shows the average accuracy for various syntactic categories, revealing negative transfer effects across all the phenomena. Notably, 'Displacement' exhibited the most significant discrepancy. This discrepancy could be attributed to several factors: first, Korean permits a greater lack of displacement compared to English. Second, the stricter word order and syntactic constraints in English may have posed an additional challenge for the En(L1)-Ko(FT-L2) model's adaptation to displacement during L2 Korean learning.

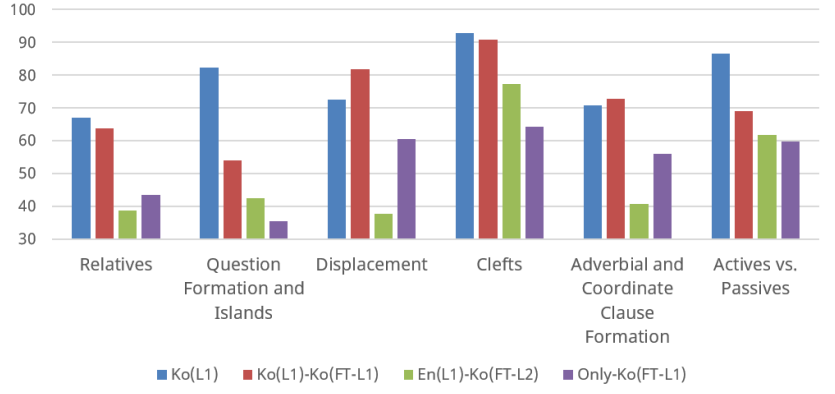


Figure 5. Each Model's Average Accuracy for Each Syntactic Phenomenon

In English, negative polarity items (NPIs), which rely on subtle semantic cues, have proven challenging for NLMs (Warstadt et al., 2020). Interestingly, our findings in Figure 6 reveal significantly high performance by the En(L1)-Ko(FT-L2) model on Korean NPI tasks, suggesting that Korean NPIs might be easier to process, thanks to the positive transfer from the learning of L1 English NPIs. However, this success might not be generalizable. Notably, the En(L1)-Ko(FT-L2) model exhibits considerably lower performance on 'Negations', 'Argument Structure', and 'Adjectives vs. Adverbs', compared to its Ko(L1) counterpart, underscoring the potential influence of L1 interference in learning the syntax-semantics of these phenomena.

Furthermore, an analysis of the average accuracy rates for the three whole categories was conducted. Figure 7 illustrates the performance contrast between the four models. In the Ko(L1) and Ko(L1)-Ko(FT-L1) NLMs, all the categories exhibited higher average accuracy, compared to Korean-fine-tuned NLMs.

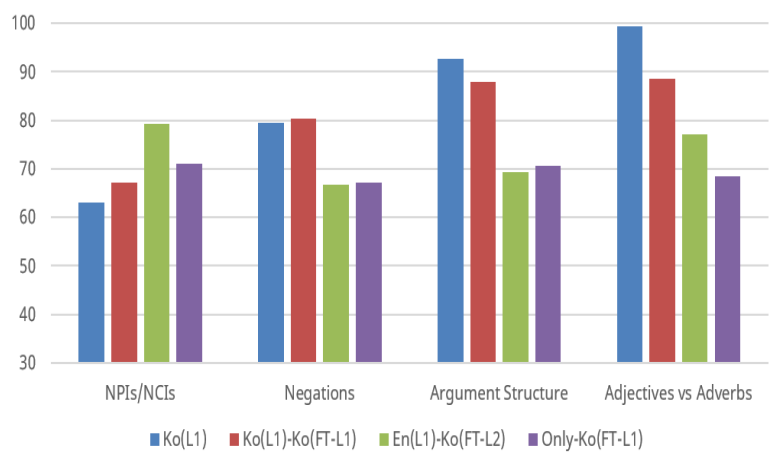


Figure 6. Each Model's Average Accuracy for Each Syntax-Semantic Phenomenon

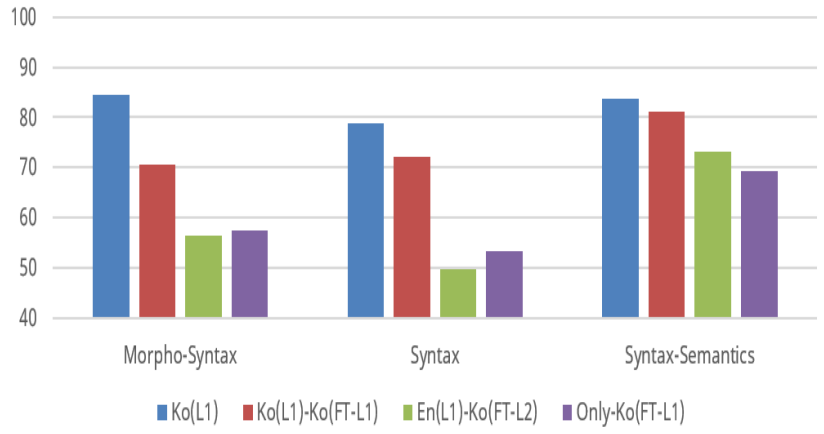


Figure 7. Each Model's Average Accuracy for the Three Whole Categories on the K-BLiMP

4.2 The Learning Trajectory for Each Test Suite on the K-BLiMP

This section explores the process of acquiring a second language, focusing on the dynamics of L2 knowledge acquisition. By using the intermediate checkpoints of our NLMs (up to 25,000 training steps), we evaluated the development of Korean grammatical knowledge over time, as shown in Figure 8. Furthermore, we examined the progress of the NLM trained exclusively on the dataset we collected for fine-tuning. This analysis helps us determine whether the En(L1)-Ko(FT-L2) NLM demonstrates genuine L1 interference effects.

Figure 8 shows the overall trajectories of grammatical abilities in the NLMs at issue. The graph represents the overall accuracy of each NLM in the progress of training steps, with the x-axis indicating the number of training steps and the y-axis indicating accuracy. Each training step corresponds to an update of the NLM's parameters using a subset of the training data.

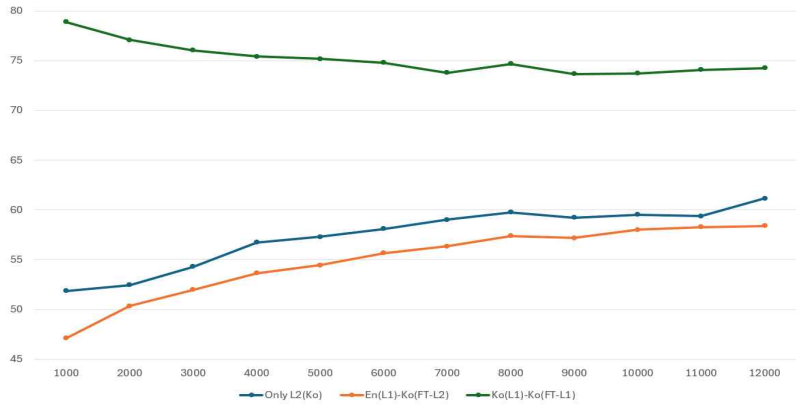


Figure 8. The Korean Grammar Learning Trajectory of Each Model

Figure 8 shows that in the initial fine-tuning stage, Ko(L1)-Ko(FT-L1) does not actively acquire new grammatical knowledge. Instead, it relies on the "already generalized" inductive bias from the pre-training dataset. As a result, Ko(L1)-Ko(FT-L1) initially performs well on the BLiMP tasks due to its knowledge from datasets in pre-training. However, as the fine-tuning dataset becomes more generalized, its performance begins to decline.

Figure 8 also shows that the Only-Ko(FT-L1) model exhibits intermediate performance, surpassing En(L1)-Ko(FT-L2) but not reaching the level of Ko(L1)-Ko(FT-L1). This observation supports the fact that En(L1)-Ko(FT-L2) initially performs poorly in processing Korean due to its English L1 transfer effects. However, the progressive increase in En(L1)-Ko(FT-L2) performance during fine-tuning suggests its capacity to gradually acquire Korean grammatical knowledge. But the significant performance gap between En(L1)-Ko(FT-L2) and Only-Ko(FT-L1) models indicates that since both models use the same dataset in the course of fine-tuning, the former's lower performance is not due to the dataset size, but due to the L1 transfer effects.

Additionally, we conducted a statistical analysis to understand the progressive increase in performance in more details, using linear regression with step as an independent variable and accuracy as a dependent variable. The results for the

overall progressive increase in Only-Ko(FT-L1), En(L1)-Ko(FT-L2), and Ko(L1)-Ko(FT-L1) are shown in Table 4. For Only-Ko(FT-L1), the linear regression model indicates a statistically significant relationship between step and accuracy, as evidenced by the F-statistic: $F(1,10)=187.6$, $p<0.001$. This suggests that accuracy is expected to increase by 0.0005 units per step. Similarly, for En(L1)-Ko(FT-L2), the linear regression model also shows a statistically significant relationship between step and accuracy, with an F-statistic of $F(1,10)=74.45$, $p<0.001$, indicating an expected increase in accuracy of 0.0009 units per step. Ko(L1)-Ko(FT-L1) registers a statistically significant relationship between step and accuracy. However, the F-statistic for this model is $F(1,10)=25.97$, $p<0.001$, showing that accuracy is expected to decrease by 0.0004 units per step.

Table 4. Overall Results of the Linear Regression Analysis for the Models

model	linear regression	F-value
Only-Ko(FT-L1)	accuracy ~ 51.8833 + 0.0005 × step	$F(1,10)=187.6$***
En(L1)-Ko(FT-L2)	accuracy ~ 48.8864 + 0.0009 × step	$F(1,10)=74.45$***
Ko(L1)-Ko(FT-L1)	accuracy ~ 77.5514 − 0.0004 × step	$F(1,10)=25.97$***

'***' $p< 0.001$ '**' $p< 0.01$ '*' $p< 0.05$

To summarize, this analysis reveals that the NLMs at issue acquire grammatical knowledge from fine-tuning data based on their L1 inductive biases. This phenomenon, referred to as transfer effects, has been investigated in this section, with a specific focus on how it evolves throughout the learning trajectory.

5. Discussion

In this paper, we investigated how the grammatical knowledge obtained in the first language (English) for the NLMs influences the learning of grammatical structures in the second language (Korean). Our findings from the Transformer-based LM, En(L1)-Ko(FT-L2), which was pre-trained in English and subsequently fine-tuned in Korean, show that the grammatical knowledge from English influenced the learning of Korean grammatical structures. Moreover, we observed that the NLM pre-trained in English exhibited negative transfer effects, leading to errors in learning Korean.

First, the accuracy of grammatical phenomena indicates that En(L1)-Ko(FT-L2), when exposed to Korean as the second language during fine-tuning, approached chance-level performance in 3 out of the 14 phenomena evaluated in the K-BLiMP test suite. In contrast, Ko(L1), only fine-tuned in Korean as its first language, achieved higher accuracy compared to En(L1)-Ko(FT-L2). Even the Only-Ko(FT-L1) model demonstrated higher accuracy than En(L1)-Ko(FT-L2). This finding can be attributed to the negative transfer from the English pre-training of the NLM.

Second, the observed differences between En(L1)-Ko(FT-L2) and Ko(L1) in each grammatical phenomenon indicate significant negative effects. En(L1)-Ko(FT-L2), fine-tuned with exposure to Korean as the second language, recorded lower accuracy across 11 of the 14 phenomena, compared to Ko(L1), fine-tuned with exposure to Korean as the first language. A parallel study by Yadavalli et al. (2023) registered similar transfer effects when learning English as the second language after L1 Japanese, using the BLiMP test to evaluate the NLMs. Their findings highlight the challenges faced by the NLMs trained in Japanese when acquiring English as the second language, demonstrating negative transfer effects. Similarly, Koo & Kim (2023) studied transfer effects involving an NLM, focusing on the instances where Korean served as the first language and English as the second language. Their approach, also using the BLiMP test rather than the K-BLiMP test, presented an alternative viewpoint to the current research. The results of Koo & Kim (2023) were consistent with those of the current study; they reveal negative transfer effects when their NLM was trained

with Korean as L1 and English as L2.

Finally, to capture the learning progression of our models, we applied the trajectory method. A detailed trajectory analysis showed that as training progressed, the En(L1)-Ko(FT-L2) model acquired grammatical features related to Korean syntax, morpho-syntax, and the syntax-semantics interface. However, its performance remained significantly inferior to that of the Ko(L1) model. This disparity demonstrated that the interference from the NLM's second language, English, posed a substantial obstacle in effectively learning Korean as the second language. Furthermore, the trajectory analysis revealed that the second language-acquiring NLM did not fully capture the grammatical features of Korean during progressive training steps, compared to Ko(L1).

Based on these observations, we conclude that negative transfer effects occur, indicating that the inductive bias embedded in the NLM's first language, English, negatively impacted its ability to learn or acquire Korean as the second language.

6. Conclusion

This paper investigated the influence of grammatical knowledge acquired in the first language (English) of the LMs on the learning of grammatical structures in the second language (Korean). Our findings highlight two main points. First, we identified negative transfer effects during the learning process from English as L1 to Korean as L2. Second, the current trajectory analysis showed that the second language acquisition model captured grammatical features of Korean, including syntax, the syntax-semantics interface, and morpho-syntax, during progressive training steps. Our study on second language learning contributes to predicting potential grammatical challenges arising from interference by the first language during the acquisition of Korean as the second language. Future research should focus on developing strategies to mitigate negative transfer effects and enhance the positive transfer of grammatical features in both human learners and NLMs.

In developing a linguistically human-like NLM to explore the issues

addressed in this study, we exercised selectivity in choosing the neural network algorithm and the NLM evaluation metric, which introduces certain limitations to our research. The findings reported are inherently tied to the specific NLM used in our experiments. Different neural network architectures or pre-training methodologies could yield varying results, meaning the conclusions drawn here may not be universally applicable to all types of NLMs. Furthermore, the metric used to assess model performance in this study may not comprehensively reflect the NLM's grammatical abilities. Alternative metrics or evaluation techniques could offer additional perspectives on the NLMs' capacity for language acquisition. These considerations will be addressed in future research.

References

Artetxe, M., Labaka, G., & Agirre, E. (2018). Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *Proceedings of the AAAI Conderence on Artificial Intelligence*, New Orleans: Hilton New Orleans Riverside.

Berzak, Y., Reichart, R., & Katz, B. (2014). Reconstructing native language typology from foreign language usage. *Proceedings of the Eighteenth Conference on Computational Language Learning,* Baltimore: Baltimore.

Blevins, T., Gonen, H., & Zettlemoyer, L. (2022). Analyzing the mono- and cross-lingual pre-training dynamics of multilingual language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* Abu Dhabi: Abu Dhabi.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of multilingual and multicultural development, 26*(1), 1-11.

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018).

Word translation without parallel data. *Proceedings of the ICLR 2018 Conference Blind Submission, Vancouver: Vancouver Convention Center.*

Deshpande, A., Talukdar, P., & Narasimhan, K. (2021). When is bert multilingual? isolating crucial ingredients for cross-lingual transfer. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle: Seattle.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis: NAACL.

Dulay, H. C., and Burt, M. K. (1974). Errors and strategies in child second language acquisition. *TESOL Quarterly, 8*(2), 129-136

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification.

Koo, K., & Kim, E. (2023). Investigating syntactic transfer in second language learning of neural language models. *The Journal of Linguistics Science, 107*, 67-81.

Oba, M.., Kuribayashi, T., Ouchi, H.., & Watanabe, T. (2023). Second language acquisition of neural language models. *Findings of the Association for Computational Linguistics: ACL 2023*, 13557−13572.

Papadimitriou, I. & Jurafsky, D. (2020). Learning music helps you read: Using transfer to study linguistic structure in language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Online.

Pérez-Mayos, L., Ballesteros, M., & Wanner, L. (2021). How much pre-training data do language models need to learn syntax?. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* Punta Cana: the Barceló Bávaro Convention Centre.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, Brussels: SQUARE.

Ri, R. & Tsuruoka, Y. (2022). Pre-training with artificial language: Studying transferable knowledge in language models. *Proceedings of ACL*, 7302-7315, Dublin: Association for Computational Linguistics.

Ruder, S., Vulić, I., & Sgard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research, 65*, 569-63.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second language acquisition modeling. In Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, & Helen Yannakoudakis (Eds.), *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications,* (pp.56–65), New Orleans: Association for Computational Linguistics.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., ... & Wei, J. (2022). Language models are multilingual chain-of-thought reasoners. *Proceedings of the ICLR 2023 poster,* Kigali: the Kigali Convention Center.

Sprouse, J. & Hornstein, N. (2013). Experimental syntax and island effects: Toward a comprehensive theory of islands. In Jon Sprouse & Norbert Hornstein (Eds.), *Experimental syntax and island effects,* (pp.1-20), Cambridge: Cambridge University Press

Tyler, T., Zhuowen, T., & Berge, B. (2022). The geometry of multilingual language model representations. *Proceedings of 2022 Conference on Empirical Methods in Natural Language Processing,* Abu Dhabi: Abu Dhabi National Exhibition Centre.

Vaswani, A., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., Łukasz, K., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach: Long Beach Convention Center.

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

Warstadt, A. & Bowman, S. R. (2023). What artificial neural networks can tell

us about human language acquisition. *Algebraic Structures in Natural Language,* 17-60.

Wu, S. & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* Hong Kong: Asia World Expo.

Wu, S., Conneau, A., Li, H., Zettlemoyer, L., & Stoyanov, V. (2019). Emerging cross-lingual structure in pre-trained language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Online: Online.

Yadavalli, A., Yadavalli, A., & Tobin, V. (2023). SLABERT talk pretty one day: Modeling second language acquisition with BERT. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics,* Toronto: the Westin Harbour Castle.

Appendix

Table 5. SKT-KoBERT's and KLUE-RoBERTa's Average Accuracy of Each Phenomenon

Phenomena	SKT-KoBERT	KLUE-RoBERTa
Clefts	75.38	97.49
Actives vs. Passives	41.61	93.96
Honorifics	49.18	38.11
Adverbial and Coordinate Clause Formation	65.55	72.24
Argument Structure	42.81	88.29
Tense/Aspect/Modal	80.70	100.00
NP Structure	33.67	97.99
NPIs/NCIs	34.90	65.77
Case Particles	59.04	94.38
Adjectives vs Adverbs	58.39	99.33
Displacement	59.04	91.16
Relatives	69.13	59.73
Question Formation and Islands	79.26	58.86
Negations	45.61	86.72
Average	**56.73**	**81.72**

Park, Myung-Kwan

Professor
Division of English
Dongguk University
30 Pildong-ro 1-gil, Jung-gu
Seoul 04620, Korea

Phone: +82-2-2260-3153
Email: parkmk@dgu.edu

Received on May 21, 2024
Revised version received on June 17, 2024
Accepted on June 28, 2024