Korean Journal of English Language and Linguistics, **Vol 24, June 2024, pp. 568-588**
DOI: 10.15738/kjell.24..202406.568

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

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

http://journal.kasell.or.kr

Investigating Grammatical Transfer in Korean-English GPT2 Language Models

Keonwoo Koo · Jaemin Lee · Myung-Kwan Park (Dongguk University)

Received: February 1, 2024
Revised: April 11, 2024
Accepted: June 19, 2024

Keonwoo Koo (co-1st author)
Ph.D. Candidate,
Department of English,
Dongguk University
Email: qjelrjsdn@naver.com

Jaemin Lee (co-1st author)
MA Candidate,
Department of English,
Dongguk University
Email: whd7987@gmail.com

Myung-Kwan Park
(corresponding author)
Professor,
Department of English,
Dongguk University
Email: parkmk@dgu.edu

ABSTRACT

Koo, Keonwoo, Jaemin Lee, and Myung-Kwan Park. 2024. Investigating grammatical transfer in Korean-English GPT2 language models. *Korean Journal of English Language and Linguistics* 24, 568-588.

With the recent success of artificial neural language model (LMs), their language acquisition has gained much attention (Futrell et al. 2019, Hu et al. 2020, Linzen et al. 2016, Warstadt et al. 2020, Wilcox et al. 2018). This paper delves into their second language (L2) acquisition, a largely unexplored area compared to their first language (L1) learning. The primary focus is on unraveling transfer effects originating from the L1's linguistic structures. By closely examining our LMs' performances on English grammar tasks, this study inspects how LMs encode abstract grammatical knowledge, particularly how pre-training biases acquired from Korean (L1) influence English (L2) performances in LMs. We present exploratory experiments where LMs were first trained on the dataset representing the initial language acquisition stage, followed by fine-tuning on the second language dataset. We analyzed cross-lingual transfer effects across diverse linguistic phenomena with the BLiMP test suite. We found that L1 pre-training did not accelerate linguistic generalization in the second language. Furthermore, our results revealed significant L1-interference, where the initial language knowledge hindered the LMs' ability to acquire and apply second language rules.

KEYWORDS

second language acquisition, neural language model, GPT-2, transfer effects, L1-interference

1. Introduction

In recent years, the transferability of artificial neural language models ((N)LMs) across different languages has garnered significant attention. Some studies have explored this phenomenon, highlighting the robust capabilities of large-scale English language models (Artetxe et al. 2018, Conneau et al. 2017, 2018, Ruder 2017, Wu and Dredze 2019, Wu et al. 2019). Notably, these models show impressive performance even when trained with a limited amount of non-English language data, suggesting a remarkable transfer of linguistic proficiency from English to other languages (Brown et al. 2020, Shi et al. 2023). This outstanding ability raises important questions about the underlying mechanisms of language transfer. Therefore, it has become crucial to evaluate these models using structured metrics that can provide deeper insights into their cross-lingual transferability. Such assessment of cross-lingual transferability has traditionally involved comprehensive metrics, such as perplexity and accuracy in downstream tasks (Blevins et al. 2022, Deshpande et al. 2022, Papadimitriou and Jurafsky 2020). However, there is much needed substantial scope of investigation into transfer effects from linguistic perspectives, including the acquisition of grammatical knowledge and the aspects of cross-lingual grammatical transfer among different languages.

This cross-linguistic transfer can be described as the influence of the first language (L1) properties on the language learner’s linguistic performance in the new, second language (L2). The interplay between the linguistic structure of an individual's L1 and the acquisition of the L2 gives rise to what is referred to as transfer effects. Within this context, cross-linguistic transfer manifests itself in either a positive or negative manner: positive transfer refers to the advantageous impact of one language in facilitating the acquisition of another, while negative transfer (or L1-interference) signifies the occurrence of errors arising from the differences between the learner's L1 and L2 languages. Note that the magnitude of negative effects tends to increase with greater dissimilarities between the two languages involved.

In the past several years, a substantial number of studies have investigated the linguistic ability of monolingual language models (LMs) (Ettinger 2020, Giulianelli et al. 2018, Gulordava et al. 2018, Lakretz et al. 2019, Linzen et al. 2016, Wu et al. 2020). These studies have primarily focused monolingual LMs on the syntactic properties of specific languages, while fewer studies have examined the grammatical knowledge of a second language in LMs. In this study, we explore the cross-lingual transferability of LMs, especially with the transformer-based model GPT-2, within the context of second language acquisition. Building upon previous research indicating that monolingual LMs perform substantially well, our research question focuses on examining how the acquisition of a first language by LMs influences the efficacy of grammar acquisition in a second language. To conduct a detailed examination of LMs' transfer effects, we design three experimental procedures: (i) pre-training the LMs with the first language; (ii) further training (fine-tuning) with the second language. (iii) After building LMs, we use the BLiMP (Warstadt et al. 2020), a benchmark of English grammatical judgment test to evaluate the LMs’ L2 grammatical generalization. This benchmark consists of 12 test suites; each corresponds to a specific linguistic phenomenon and falls into one of four linguistic categories: syntactico-morphology, syntax, semantics, and syntax-semantics interface.

To achieve the goal of this paper, we use Korean as the first language and employ English as the second language. We predict that transfer from Korean to English becomes difficult considering multiple factors: word order difference (SOV vs SVO), linguistic distance (Chiswick and Miller 2003, Grimes and Grimes 2002), and other things affecting learning difficulty[1].

[1] According to Foreign Language Training Institute in https://www.state.gov/foreign-language-training/, in terms of the

This paper is organized as follows. In section 2, we present previous research concerning the second language acquisition of LMs. Section 3 outlines the experimental methodology, datasets, and LMs used in this study. In section 4, we show our findings with their implications. The section 5 addresses the general discussion. The last section involves concluding remarks.

2. Previous Work

The acquisition of a second language has been a long-standing area of investigation within applied linguistics, psycholinguistics, and pedagogy (Ellis 2010, Hatch 1983, Krashen 1981). These disciplines have yielded numerous hypotheses and theories on human language learning, such as the influential Input Hypothesis (Krashen 1977). Parallel to this human-centric research, a separate line of inquiry emerged in the 1980s exploring the potential of neural models for language acquisition. Driven by the question of whether language can be acquired without innate knowledge (Pinker and Prince 1988, Rumelhart and McClelland 1986), initial investigations relied on simple neural networks. With the subsequent development of Neural NLP (Manning 2015), a renewed interest has arisen in revisiting these classic questions posed by cognitive science (Kirov and Cotterell 2018). A key trend within this field is the increasing focus on probing the linguistic knowledge embedded within neural language models (Linzen et al. 2016, Warstadt and Bowman 2022).

Furthermore, the phenomenon of language transfer in NLP models is actively investigated from both engineering and linguistic-scientific perspectives. In the engineering domain, researchers aim to mitigate the English-centric bias inherent in many NLP techniques by developing models capable of handling multiple languages with greater proficiency (Conneau et al. 2020, Dong et al. 2015, Lample and Conneau 2019). Meanwhile, from a linguistic-scientific standpoint, the mechanisms and linguistic properties underlying LMs' language transfer abilities are being explored (Blevins et al. 2022, Chang et al. 2022, Pérez-Mayos et al. 2021). Notably, this exploration extends beyond the transfer between natural languages, venturing into the realm of artificial languages (Papadimitriou and Jurafsky 2020, Ri and Tsuruoka 2022). One of the primary motivations behind such analyses is to quantify the transferable universals that lie behind both human and artificial languages.

In this realm, cross-lingual transfer has received considerable attention in NLP research (Artetxe et al. 2018, Conneau et al. 2017, 2018). However, most of this research has more focused on practical implications such as which tokenizer can optimize cross-lingual transfer and did not give much attention to the kind of sequential transfer relationships that arise in human second language acquisition or learning.

Meanwhile, in recent years, there have been a significant interest in investigating patterns of positive and negative transfer of LMs (Oba et al. 2023, Papadimmitriou and Jurasky 2020, Yadavalli et al. 2023). They applied various sorts of methodology and metrics to demonstrate valuable implications.

Papadimmitriou and Jurasky (2020) as well as Yadavalli et al. (2023) employed the Test for Inductive Bias via Language Model Transfer (TILT) method in their studies. This approach enables the freezing of pre-trained data parameters in LMs, excluding the word embeddings. Subsequently, fine-tuning is conducted using the second language, affecting only the embedding layer. Through this method, they sought to discern instances where LMs, when learning only vocabulary and not the structure of sentences, exhibit positive transfer or interference effects.

learning difficulty level, Korean is among “super-hard languages”. Note that the difficulty levels provided in this site exclusively indicate the challenges associated with transferring from English to a specific language. In our study, we tentatively assume the symmetry between the source and target languages concerning the alleged learning difficulty.

Furthermore, building upon the work of Huebner et al. (2021), which demonstrated the superior linguistic abilities of Child-Directed Speech (CDS; Huebner and Willits 2021) for even models that are down-sized (100M-tokens), compared to the Wikipedia dataset. Yadavalli et al. (2023) further explored the efficacy of the age-ordered (AO) CDS dataset in CHILDES specifically designed for L1 training. Notably, they found that positive transfer, where L1 knowledge facilitates L2 acquisition, was more prevalent than negative transfer in their models, mirroring observations in human English Second Language (ESL) learners.

Building upon previous research on cross-linguistic transfer in LMs, this study investigates the extent to which GPT-2 L2 language models exhibit such effects from Korean (assumed to be an L1) to English (assumed to be an L2). Leveraging the BLiMP test suite's comprehensive assessment of syntactico-morphology, syntax, semantics, and syntax-semantics interactions, we aim to provide a more nuanced understanding of how L2 LMs handle the complexities of L2 acquisition. By analyzing their performances across diverse linguistic tasks and analyzing transfer patterns, this study aspires to investigate cross-lingual knowledge transfer in LMs and shed lights on the specific challenges and advantages encountered during L2 learning. Specifically, we will investigate the following research questions:

(i) How do neural language models pre-trained on typologically distinct languages, such as Korean and English, differ in their ability to acquire and generalize second language grammar rules?

(ii) What specific challenges do language models face when transferring syntactic structures from a Subject-Object-Verb (SOV) language like Korean to a Subject-Verb-Object (SVO) language like English?

(iii) How does the complexity of linguistic phenomena—such as anaphor agreement, quantifiers, and filler-gap dependencies—impact the performance of L2 language models pre-trained on different L1 datasets?

3. Experiment

3.1 Overall Procedures

This section outlines the comprehensive experimental methodology utilized in the study. The process is divided into several critical phases to examine cross-lingual transfer effects and the acquisition of grammatical knowledge by LMs in a structured manner.

We first pretrain the models on large-scale datasets representing the L1. Specifically, we use Korean Wikipedia (Ko-Wiki) for the Korean model and English Wikipedia (En-Wiki) for the English model. This phase simulates the natural language exposure during early language acquisition in humans.

Following pretraining, the models undergo a fine-tuning process on datasets representing second L2 learning. Two datasets are employed: the K-English Textbook dataset, which includes English learning materials used by Korean students, and the Adult-Directed Speech (ADS) corpus, reflecting naturalistic English language exposure.

The models are evaluated using the BLiMP test suite to assess their grammaticality judgment capabilities. The BLiMP suite includes 12 test suites, each corresponding to different linguistic phenomena, categorized into syntactico-morphology, syntax, semantics, and syntax-semantics interface.

3.2 Datasets

3.2.1 L1 datasets

Studies have shown that by the age of ten, humans have acquired an estimated one hundred million words in their native language (Warstadt and Bowman 2022). To replicate this human language acquisition in our language models, we gathered two datasets amount of one hundred million words each: (i) one from the Korean Wikipedia (henceforth, Ko-Wiki) for the Korean model, (ii) one from the English Wikipedia (henceforth, En-Wiki) for the English model. These datasets[2] will serve as the first-language training data for our comparative study of language acquisition in our models.

3.2.2 L2 datasets

For the second language, we selected two distinct datasets to investigate the impact of different learning materials on cross-linguistic transfer: K-English Textbook and ADS (Adult-Direct Speech; Yadavalli et al. 2023) corpus. The K-English Textbook dataset comprises approximately 2.5 million tokens Korean L2 learning materials, specifically extracted from EBS-CSAT English Prep books (2016-2018) and English textbooks used in middle and high schools (2001-2002, 2009-2010, 2015-16) (Koo et al. 2022, 2023). This curated dataset reflects contemporary Korean L2 learning environments and provides a structured foundation for the models. This dataset is intended to mimic Korean human learners learning English as a second language.

The ADS corpus encompassing around 900,000 tokens, as described by Yadavalli et al. (2023), provides a valuable resource of authentic English speech for fine-tuning. This diverse and naturalistic data composition will fulfill our expectation that LMs will not only learn the natural expressions commonly used by native speakers but also develop grammatically accurate complex sentences, aligning with our objective of examining the influence of conversational input on second language acquisition or learning. The current study thus employed the two types of dataset—the L2 English learner and the ADS datasets—to determine which is most beneficial for acquiring English as a second language, particularly for LMs.

In addition, we also fine-tuned the English pretrained model with L2 learner English data, which is akin to native English speakers learning English through L2 textbooks, an unusual scenario. Native speakers might experience language attrition, as simplified and potentially incorrect structures in L2 materials could interfere with their existing knowledge. In tandem, this approach will highlight the limitations and challenges that L2 learner English data pose for LMs already proficient in English.

3.2.3 Test suite: BLiMP benchmark[3]

We use the BLiMP, a benchmark of English grammatical judgment test to evaluate our models' L2 linguistic generalization.[4] The dataset consists of 67 minimal pair types in English, each consisting of 1,000 pairs,

[2] These datasets are gathered from https://github.com/toizzy/wiki-corpus-creator.

[3] https://github.com/nyu-mll/jiant/tree/blimp-and-npi/scripts/blimp.

[4] Prior to the BLiMP test suit, the Corpus of Linguistic Acceptability (CoLA) counterpart was compiled by Warstadt et al. (2019) to track advances in the sensitivity of reusable sentence encoding models to acceptability. Refer to Warstadt et al. (2020, p. 379) for its limitations.

organized into twelve broad suites; each corresponds to a specific linguistic phenomenon. These twelve broad phenomena were selected from syntax textbooks such as Sag et al. (2003), Adger (2003), and Sportiche et al. (2013). Afterward, the minimal pairs were generated artificially based on the data using the linguist-crafted grammar templates. This can also fall into one of four linguistic categories: syntactico-morphology, syntax, semantics, and syntax-semantics interface. Overall, each test suite has 1,000 minimal sentence pairs, and each pair consists of grammatically acceptable and unacceptable as shown in Table 1. This BLiMP benchmark was validated with crowd-sourced human judgments. (More details can be found in Warstadt et al. 2020.)

Table 1. Examples of the BLiMP Test Suite (Warstadt et al. 2020)

12 Phenomena	Acceptable Example	Unacceptable Example
Anaphor Agreement	Susan could listen to herself.	Susan could listen to himself.
Argument Structure	Most banks have praised Raymond.	Most jackets have praised Raymond.
Binding	Angela was not thinking that she likes Susan.	Angela was not thinking that herself likes Susan.
Control/Raising	There is soon to be a cat existing.	There is willing to be a cat existing.
Determiner Noun AGR.	Craig had cared for that dancer.	Craig had cared for those dancer.
Ellipsis	That book bored many troubled sons and Theresa bored few.	That book bored many sons and Theresa bored few troubled.
Filler-gap Dependency	Susan knew that man that April remembered.	Susan knew who April remembered that man.
Irregular Forms	The mushroom went bad.	The mushroom gone bad.
Island Effects	Which teenagers had Tama hired and Grace fired?	which had Tama hired teenagers and Grace fired?
NPI Licensing	Tama really exited those mountains.	Tama ever exited those mountains.
Quantifiers	No girl attacked fewer than two waiters	No girl attacked at most two waiters.
Subject-verb AGR.	The child is not attacking Becky.	The children is not attacking Becky.

3.3 Experimental Setup

In this study, we use Transformer-based LM, GPT-2[5] (Generative Pre-trained Transformer 2; OpenAI). GPT-2 is a large-scale language model using the Transformer architecture (Vaswani et al. 2017). Notably, Warstadt et al. (2020) demonstrated that GPT-2 achieved superior results on the BLiMP test suite compared to other models.[6] The choice of GPT-2 aligns with our research focus on cross-linguistic transfer effects, as its ability to analyze grammatical structures makes it well-suited to explore how grammatical knowledge from Korean can be transferred to English in neural language models.[7]

[5] We built our models from https://github.com/huggingface/transformers/tree/main/examples/pytorch/language-modeling.

[6] We adopt the decoder-based language model GPT-2, which is known to specialize in sentence production rather than in sentence comprehension. But it is not precluded from carrying out comprehension tasks like grammaticality judgment. One of its strength lies in identifying unnatural word flow or nonsensical phrasings. This can help with flagging statistically unlikely word orders that might indicate grammatical errors. As mentioned in the text, since its release in 2019, Warstadt et al. (2020) demonstrated that GPT-2 performed better on grammatical judgment tasks like the BLiMP tests compared to other then current models.

[7] For the purpose of the study in this paper, we decide on GPT-2, the second in the foundational series of GPT models. Since it was publicly available in 2019, two or more recent versions have been released: GPT-3 in 2020 (or GPT-Neo in 2021) and GPT-4 in 2023. As an anonymous reviewer of this paper points out, we should have adopted the two recent versions that can make far superior generalizations/learning than GPT-2. But as pointed out in the text, we also had in mind the development of

Pretraining in neural language models is analogous to human first language acquisition as both involve initial exposure to large amounts of structured language input to learn fundamental linguistic features. In humans, this is akin to the rich, repetitive child-directed speech that forms the basis of first language learning. Fine-tuning after pretraining is similar to human second language acquisition, where the existing language knowledge (L1) is adapted to new contexts (L2), experiencing both positive and negative transfer effects. This process in models mirrors how humans use their first language framework to learn a second language, often with explicit instruction and practice.

Specifically, to model second language acquisition, we utilize fine-tuning, a method of training a pre-existing model further. This process parallels how humans acquire a second language, aligning with previous research on second language modeling. While the TILT approach (as described in section 2) limits the full acquisition of second language learning by, for example, restricting training only to the embedding layer, it does not fully replicate the authentic second language acquisition process. Therefore, in this study, we consider fine-tuning as part of the second language learning process through natural language exposure but avoid using the freezing method.

Thus, we build two L1-models with datasets mentioned in section 3.1: first, we built the models which are only pre-trained with En-Wiki (henceforth. En-Pre) and Ko-Wiki (henceforth, Ko-Pre). After building the En-Pre and Ko-Pre L1-models, we fine-tuned these models with K-English textbooks (henceforth, En-Textbook, Ko-Textbook) and ADS dataset (henceforth, En-ADS, Ko-ADS). Importantly, note that our second language models underwent fine-tuning ten epochs with different random seeds, and the reported scores were averaged across ten iterations.

4. Results and Implications

For thorough examinations, we present not only the results obtained from the model evaluations but also provide their implications in this section. Note that all the accuracy scores reported in tables and figures are presented as percentages for clarity.

4.1 Overall Results

All the models' accuracy is evaluated through their performance on the BLiMP test suite. Each minimal pair is considered, and the accuracy is determined by whether it assigns a higher probability to the grammatically correct sentence compared to the grammatically incorrect one.[8] The outcomes for all the models are illustrated in Figure

a linguistically more 'human-like' language model, thus training it with 100 million words that 10-year-old human learners are allegedly exposed to rather than 300 billion words that GPT-3 are trained with. We also took into account the finding that more recent versions of LMs can confront overfitting when they learn patterns from the training data so well that they perform poorly on new, unseen data because they have essentially memorized the training data rather than learned to generalize from it (See Radford et al. (2019) and other works following it). Though for this paper we cannot adopt them for the reasons mentioned above, we hope to carry out the same experiment using more recent versions of LMs in the near future, investigating the transfer effects in the state-of-art ones from NLP perspectives.

[8] The LMs do not allocate probabilities that sum to 100% for each pair; instead, both grammatically correct and incorrect sentences are independently assigned probabilities, and the sentence with the higher probability is considered the grammatical one.

As an anonymous reviewer of this paper pointed out, there might be a data loss issue since the higher probability of a

1. Additionally, Table 2 provides the average accuracy of each model on the BLiMP test. Detailed results can be found in the Appendix.

Table 2. Average Accuracy of Our Models on the BLiMP Test

	En-Pre	En-Textbook	En-ADS	Ko-Pre	Ko-Textbook	Ko-ADS
Avg accuracy	**65.04**	62.06 (0.27)	63.19 (0.36)	47.07	56.10 (0.37)	55.99 (0.39)

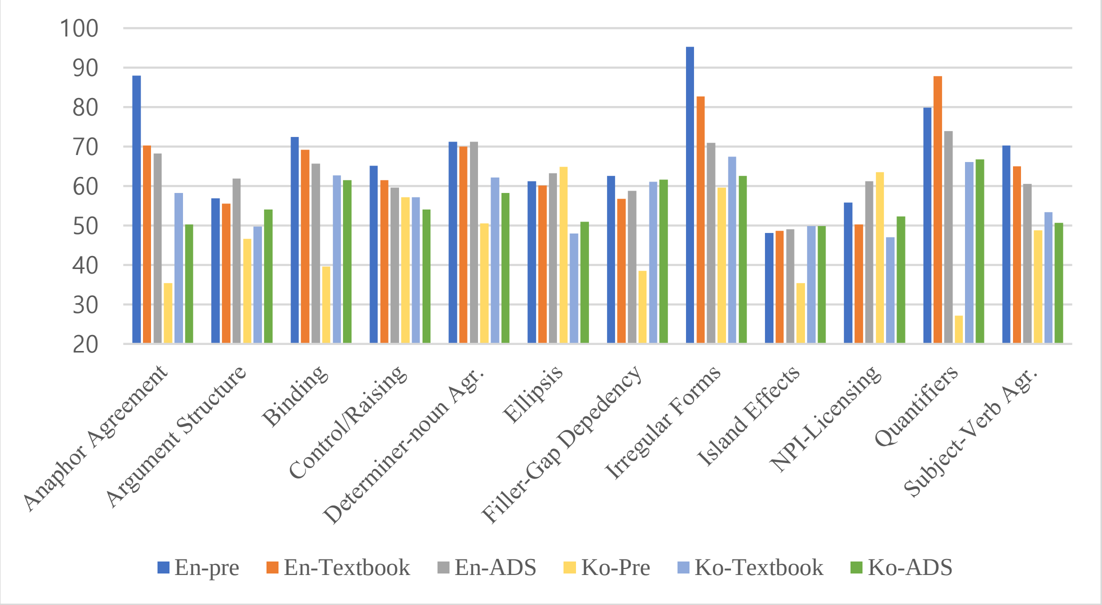


Figure 1. Overall Average Scores of the L2 Phenomena in the BLiMP

The model only pre-trained on En-Wiki, En-Pre, demonstrated the highest level of accuracy (65.04), while the model trained solely on Ko-Wiki exhibited the lowest accuracy (47.07). Despite the En-Pre model not fine-tuned on other datasets, it achieved the highest performance in 9 out of 12 phenomena tested in the BLiMP test suite (Anaphor agreement, Argument structure, Binding, Control/Raising, Determiner-noun agreement, Filler-gap dependency, Irregular forms, Quantifiers, Subject-verb agreement), as shown in Figure 1. On the other hand, the Ko-Pre model hovered around the chance level.

Furthermore, the noticed reduction in performance for the model fine-tuned with K-English textbooks was 5.96 points (En-Textbook minus Ko-Textbook). The similar reduction was observed in ADS as well with 7.2 points (En-ADS minus Ko-ADS in Table 2).

grammatical sentence can range from 50.1 to 99.9. However, given the evaluation dataset's size of 134,000 sentences, maintaining such granular scores across extensive datasets could result in significant computational overhead and complexity. In some scenarios, the marginal benefit of ensuring such precision might not justify the added complexity and computational burden. Therefore, we deemed it inefficient to control all the model's probability scores for 134,000 sentences across the six models. Although this paper does not resolve this issue, we plan to use more advanced metrics, such as Log-Loss (Cross-Entropy Loss) (cf. Niu and Penn 2020), in future research.

4.2 Comparison Between En-LMs and Ko-LMs

Since we are interested in capturing the cross-lingual transfer effects between Korean and English, we closely compare En-Textbook versus Ko-Textbook, as well as En-ADS versus Ko-ADS based on the average accuracy of each 12 phenomenon.

In Figure 2, the two models were fine-tuned with K-English textbook, and it was observed that the Ko-Textbook model exhibited sub-par performances across all the phenomena except for Filler-gap Dependency. Additionally, as Figure 3 demonstrates, the most significant performance drop was observed in Quantifiers, decreasing by 21.7. Notable decreases were also evident in Irregular Forms (15.27 points), ellipsis (12.2 points), Anaphor Agreement (11.99 points), and Subject-verb Agreement (11.61 points) respectively (Details are illustrated in Appendix). From these aspects, we can infer that the Ko-Textbook model demonstrated negative transfer effects.

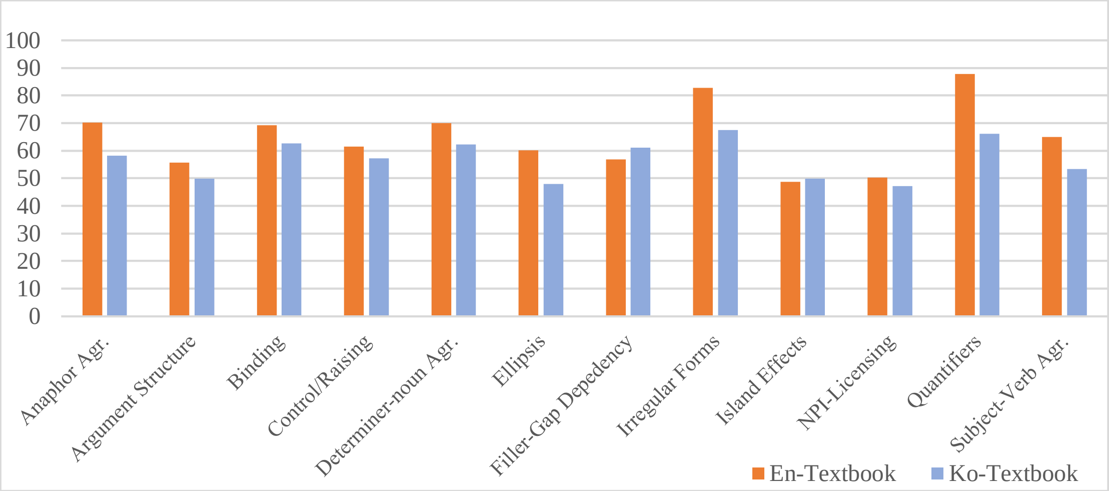


Figure 2. Overall Average Scores of En-Textbook and Ko-Textbook Models for the BLiMP

To confirm language transfer effects, we performed a t-test using the accuracy of each phenomenon in En-Textbook and Ko-Textbook. The results, summarized in Table 3, show significant differences between En-Textbook and Ko-Textbook in each phenomenon, indicating negative transfer effects (all *p*'s < 0.01).

Table 3. The Results of t-test Between the En-Textbook and Ko-Textbook Model

Phenomena	t-value	df	*p*-value
Anaphor Agr.	9.1220	9	***
Argument Structure	20.3022	9	***
Binding	22.4764	9	***
Control/Raising	15.4257	9	***
Determiner-noun Agr.	42.1325	9	***
Ellipsis	21.3546	9	***
Filler-gap Dependency	-11.7787	9	***
Irregular Forms	12.9857	9	***
Island Effects	-4.9628	9	***
NPI Licensing	4.2380	9	**
Quantifiers	30.9682	9	***
Subject-verb Agr.	65.1617	9	***

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

As shown in Figure 3, the two models were fine-tuned to the ADS corpus, exhibited similar negative effects to those in Figure 2. However, the big difference between En-ADS and Ko-ADS was that Ko-ADS performed worse particularly in agreement phenomena (e.g. Anaphor Agr, Determiner-noun Agr, Subject-verb Agr.) compared to En-ADS model (Details are illustrated in Appendix). The observed negative transfer effects can be accounted for by the remarkable difference in syntactic agreement relation between Korean and English: unlike the latter, the former lacks the syntactico-morphological agreement relation. The increased flexibility of word order in Korean sentences involving scrambling, compared to the stricter word order restriction in English, might also have challenged the LM's ability to adapt the learning of English as an L2.

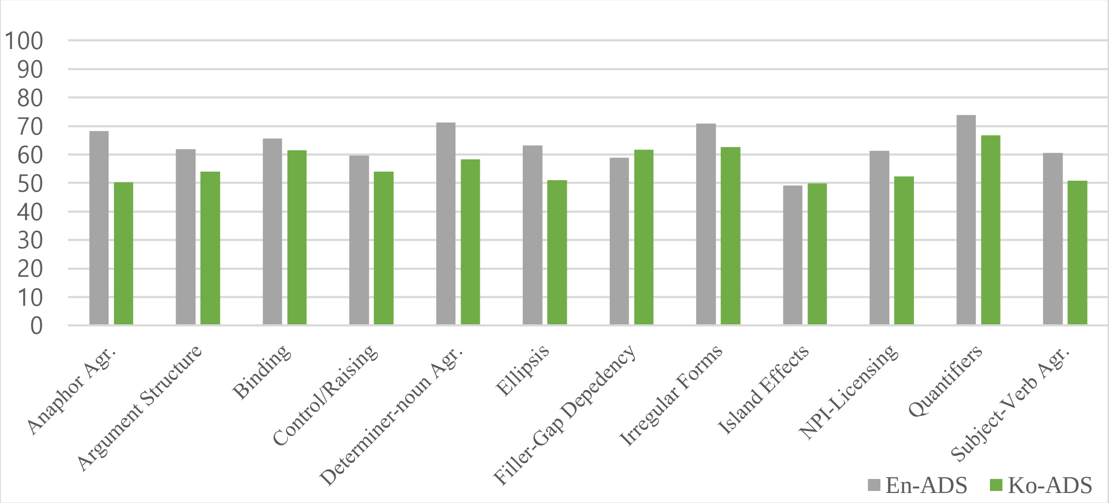


Figure 3. Overall Average Scores of En-ADS and Ko-ADS Models for the BLiMP

We performed the t-test again using the accuracy of each phenomenon in En-ADS and Ko-ADS, with the results summarized in Table 4. The differences between En-ADS and Ko-ADS in each phenomenon reveal significant negative transfer effects (p < 0.01), except for the Island phenomenon. The Island phenomenon will be discussed in the next section (See section 4.3).

Table 4. The Results of t-test Between the En-ADS and Ko-ADS Model

Phenomena	**t-value**	**df**	***p*-value**
Anaphor Agr.	-23.8826	9	***
Argument Structure	-43.2187	9	***
Binding	-10.2968	9	***
Control/Raising	-12.6142	9	***
Determiner-noun Agr.	-90.2288	9	***
Ellipsis	-24.5491	9	***
Filler-gap Dependency	6.8539	9	***
Irregular Forms	-6.4305	9	**
Island Effects	1.4522	9	*0.1803*
NPI Licensing	-12.2705	9	***
Quantifiers	-8.4460	9	***
Subject-verb Agr.	-40.0343	9	***

‘***’ p<0.001 ‘**’ p<0.01 ‘*’ p<0.05

To sum up this sub-section, the results of the twelve specific linguistic phenomena between En LMs and Ko LMs indicate that Korean grammatical features negatively impact LMs when learning English as a second language. This finding supports the prediction made above that the structural dissimilarities between the L1 (Korean) and L2 (English) adversely affect the models' performances.

4.3 Transfer Effects Categorized into Four Categories

Since the BLiMP test suite covers various aspects of linguistic features, we categorized the 12 phenomena into 4 (syntactico-morphology, syntax, semantics, syntax-semantics interface), to specifically investigate cross-lingual transfer effects of our models. The breakdowns are shown in Table 5.

Table 5. 12 Phenomena divided into Four Category

Syntactico-Morphology	Anaphor Agreement
	Determiner-noun Agreement
	Irregular Forms
	Subject-verb Agreement
Syntax	Argument Structure
	Ellipsis
	Filler-gap Dependency
	Island Effects
Semantics	NPI Licensing
	Quantifiers
Syntax-Semantics Interface	Binding
	Control/Raising

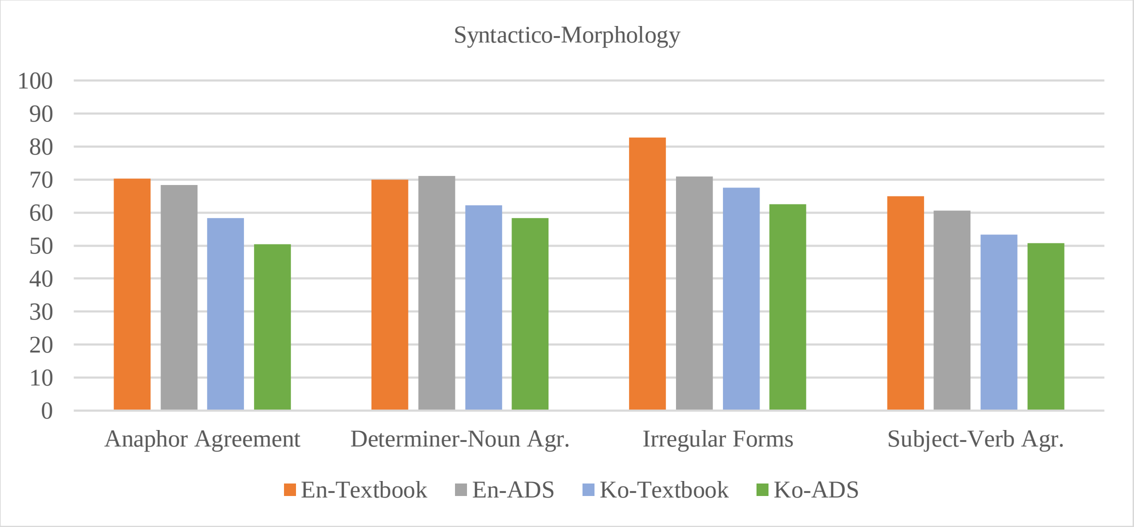


Figure 4. Average Accuracies of the Models in Syntactico-Morphology

Despite the relative simplicity of Determiner-noun and Subject-verb Agreement, the phenomena governed by simple syntactico-morphological rules (Yadavalli et al., 2023), Figure 4 reveals that the Ko-LMs (i.e., Ko-Text and Ko–ADS) performed significantly worse than the En-LMs on syntactico-morphology tasks. This suggests that, as pointed out above, the initial acquisition of Korean syntactico-morphological rules by the Ko-LMs is most likely hindering their ability to learn English rules, where syntactico-morphology can exhibit more complexity or variation. Several linguistic dissimilarities between Korean and English syntactico-morphology may have contributed to this negative transfer. For instance, unlike English pronouns, which have distinct forms in gender and number (e.g., *he/she/they*), Korean pronouns like *ku* can be used as gender-neutral. This difference in pronoun systems may have impeded the Ko-LMs' ability to generalize agreement rules from Korean to English. Moreover, in English, subject-verb agreement manifests itself through changes in verb form depending on the subject's number (singular or plural). In contrast, Korean verbs do not allow variance in form regardless of the subject's number, retaining a single form for both singular and plural subjects. Based on such linguistic dissimilarities, these observations suggest that both Ko-Textbook and Ko-ADS patterns in negative transfer effects are indeed influenced by their initial acquisition of Korean syntactico-morphological rules and could pose challenges for the Ko-LMs while learning English as an L2.

We now move to Figure 5, which shows the average accuracy scores of syntax syntactic phenomena. We observe that negative transfer effects occurred in Ellipsis and Argument Structure. The interesting point is that in Filler-gap Dependency and Island Effects, the Ko-ADS performed slightly better than the Ko-Textbook. This indicates that since the ADS is authentic English, it produced the more positive transfer effect than the Ko-Textbook; the ADS which has a characteristic of adult-directed-speech is more appropriate than the English textbooks to train the Korean-English L2 LM. However, a caveat is in order: while the Korean-English L2 LM's slightly better performance on these two specific downstream tasks suggests some potential for positive transfer, it is premature to conclude a definitive overall transfer effect based solely on this limited evidence.

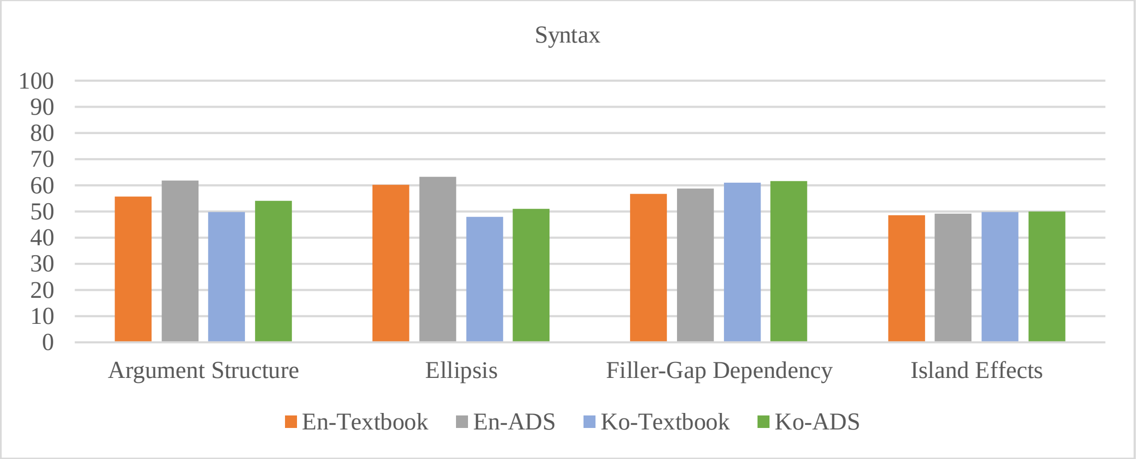


Figure 5. Average Accuracies of the Models in Syntax

Now turning to the semantic phenomena, negative polarity item (NPI) licensing has been shown to be particularly challenging for all the language models due to its reliance on semantics cues (Warstadt et al., 2020). This aligns with our findings in Figure 6, where even English-based LMs fine-tuned on the two different datasets exhibited limited performances on the task involving NPIs. Notably, the average accuracy of En-LMs on the NPI task was 50.2 (En-Textbook) and 61.2 (En-ADS), suggesting that acquiring semantic knowledge related to NPIs remains a significant hurdle for the LMs. Furthermore, the Ko-LMs displayed even lower performances (47.1; Ko-Textbook, 52.3; Ko-ADS) than their En-LM counterparts, highlighting the L1-interference in the acquisition of English semantics.

However, in the domain of Quantifiers, the En-textbook and En-ADS models achieved significantly higher performances on the task involving quantifiers. Notably, their accuracy on the quantifier task reached 87.7 and

73.9, indicating a successful acquisition of grammatical knowledge in this domain. By contrast, the Ko-Textbook and Ko-ADS models exhibited negative transfer effects, with their performances on the quantifier task falling below those of the En-LMs. This also suggests that the semantic differences in quantifier systems between Korean and English may have hampered transferability and led to L1-interference during fine-tuning.

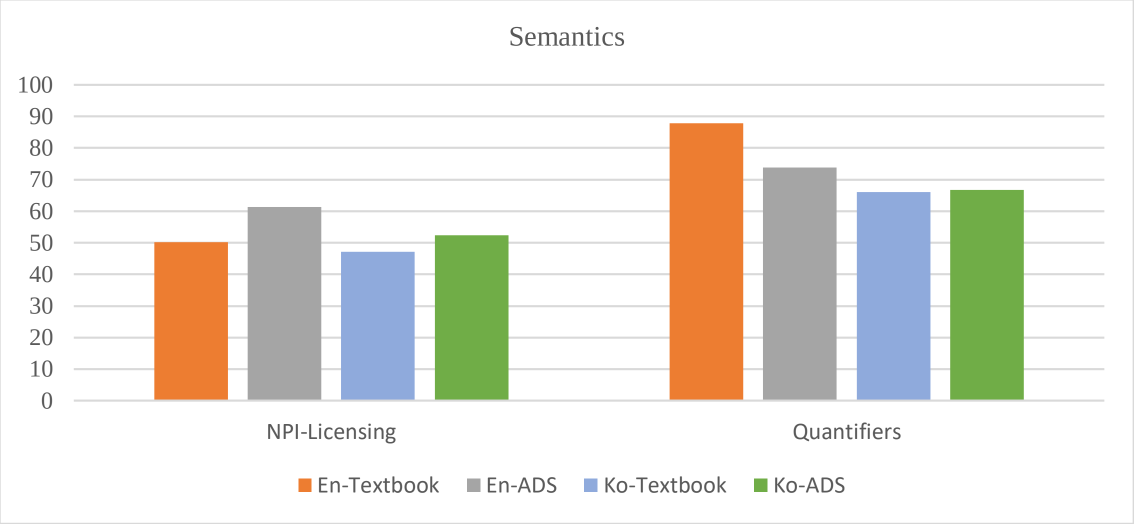


Figure 6. Average Accuracies of the Models in Semantics

Finally, Figure 7 shows the average accuracy of our models on the tasks involving syntax-semantics interfaces, phenomena requiring both syntax and semantics and their interface knowledge to perform them well (Sprouse and Hornstein 2013). Notably, the LMs pre-trained on Korean exhibited lower accuracy compared to their counterparts pre-trained on English; this suggests that the Ko-LMs' initial acquisition of Korean linguistic knowledge may have interfered with their ability to apply syntax-semantics interface rules and handle binding/control dependencies in the L2 English context.

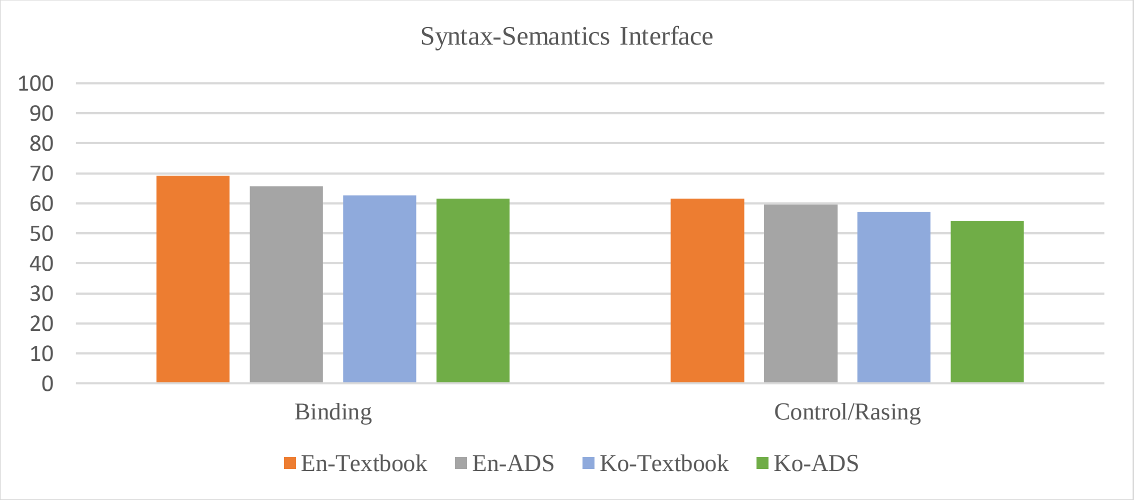


Figure 7. Average Accuracies of the Models in Syntax-Semantics Interface

5. General Discussion

The findings of the current research illuminate several crucial aspects of linguistic interference in transformer-based L2 LMs pre-trained on Korean and subsequently fine-tuned on English. The observed challenges in syntactico-morphology, semantics, syntax, and their interface as summarized in Figure 8 and Table 6 reflect inherent discrepancies between Korean and English linguistic structures. These findings contribute to our understanding of second language acquisition in neural models and highlight the areas requiring further attention for improved multilingual model training.

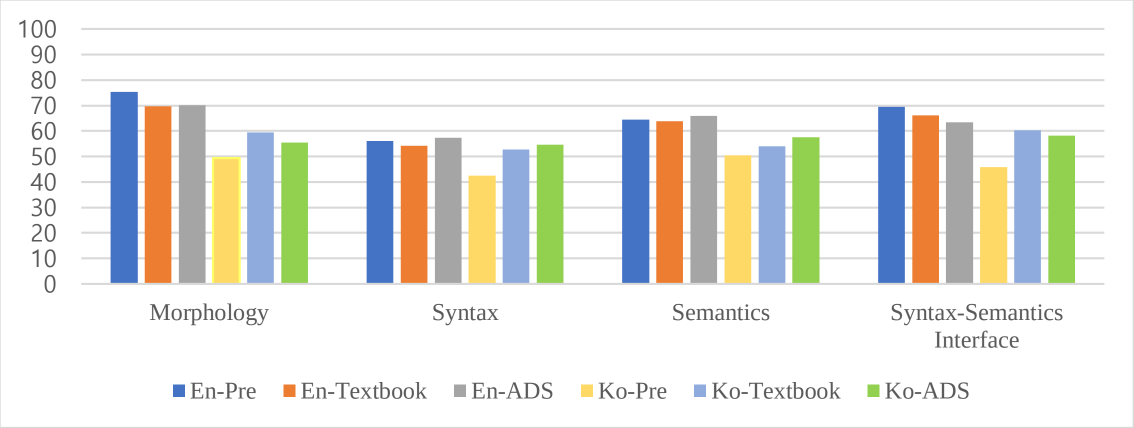


Figure 8. Four Categorized Overall Accuracies of the Models for the BLiMP Test

Table 6. Average Accuracy of the Models of Four Category

	En-Pre	En-L2	En-ADS	Ko-pre	Ko-L2	Ko-ADS
Morphology	**75.4**	69.7	70.1	49.3	59.4	55.3
Syntax	**56.0**	54.1	57.25	42.4	52.7	54.6
Semantics	64.5	63.8	**65.8**	50.3	54.0	57.55
Syntax-Semantics Interface	**69.4**	66.0	63.3	45.7	60.2	58.2

First, the Korean-English L2 LMs exhibited significant difficulty in adapting to English syntactico-morphology, particularly in aspects where Korean and English diverge sharply. For instance, the models struggled with Anaphor Agreement, Determiner-noun Agreement, and Subject-verb Agreement. English pronouns require syntactico-morphological agreement in gender and number, a feature less prominent in Korean pronouns, which are often gender-neutral and context-dependent. Furthermore, the absence of a parallel item in Korean corresponding to English articles potentially complicates the learning of Determiner-noun Agreement for the Korean-English L2 LMs. Additionally, Korean's lack of subject-verb agreement poses a substantial challenge in adapting to English's strict subject-verb agreement rules.

Second, semantics, particularly in the use of NPIs and quantifiers, presented another area of difficulty. The Korean-English L2 LMs' lower performance in the English NPI task may have stemmed from the different semantic and syntactic/distributional applications of NPIs in Korean. Likewise, the disparities in quantifier usage between the two languages may have hindered the effective transfer of semantic knowledge related to quantification.

Third, in the domain of syntax, while the interference was less pronounced, notable differences were observed in the overall score (as in Table 4). The divergence in typical sentence structure (SOV in Korean vs. SVO in English and word order permutation/scrambling) may have caused occasional word order errors in the Korean-English L2 LMs.

Lastly, we also revealed challenges in syntax-semantics interfaces, particularly in binding and control/raising constructions. Korean's flexible pronoun references contrast with the stricter rules in English reflexives and reciprocals, potentially leading to the incorrect identifications of pronoun references in the Korean-English L2 LMs. Differences in control and raising structures between the two languages may also have led to errors in identifying the correct antecedents in the English constructions at hand.

Closing this section, one may ask whether it is really true that the transfer effects reported in this paper are attributed to the relatively small size of the English textbooks used for fine-tuning the pre-trained Korean LM. To answer this question, we examine the learning trajectories of the Ko-textbook LM and the LM only trained on the K-English textbooks, as in Figure 9. We note that when the training step progresses over 6,000, the Textbook-Only LM performs better than the Ko-Textbook LM. Since there is a significant difference between the two LMs at the training step of 10,000, this clearly shows that the weak performance of the Ko-Textbook LM stems from the grammatical inference of its first language, Korean, rather than from the size of the fine-tuning dataset.

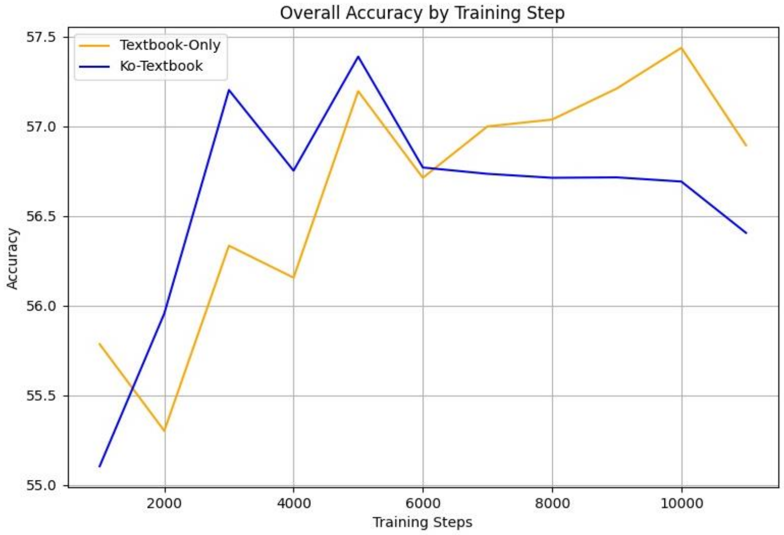


Figure 9. The Learning Trajectories of the Ko-Textbook and the Textbook-Only LMs

6. Concluding Remarks

This paper has explored the second language acquisition of neural LMs, focusing specifically on their transfer of grammatical knowledge between Korean and English. To this aim, we have trained bilingual LMs under a similar scenario to the human L2 acquisition and then analyzed their cross-lingual transfer using the BLiMP test suite. Our experiments have demonstrated that the Korean-English L2 LMs significantly lag behind than the LMs pre-trained on English in acquiring L2 English grammar.

These findings highlight the nuanced complexities involved in adapting the LM pre-trained on Korean to the second language of English, since the two languages entertain different grammatical systems. The marked differences in syntactico-morphology, semantics, syntax, and their interface between Korean and English necessitate tailored approaches in human L2 acquisition as well as model training. This research underscores the importance of considering typological differences and structural intricacies in human L2 teaching as well as the development of effective and accurate L2 LMs. Future research should focus on developing strategies to mitigate negative transfer effects and enhance the positive transfer of linguistic features in both human learners and neural LMs.

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

References

Adger, D. 2003. *Core Syntax: A Minimalist Approach.* Oxford: Oxford University Press.

Artetxe, M., G. Labaka and E. Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1.

Blevins, T., H. Gonen and L. Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3575–3590.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal and D. Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33, 1877-1901.

Chang, T. A., Z. Tu and B. K. Bergen. 2022. The geometry of multilingual language model representations. *arXiv preprint arXiv:2205.10964*.

Chiswick, B. R. and P. W. Miller. 2003. A test of the critical period hypothesis for language learning. *Journal of Multilingual and Multicultural Development 29*(1), 16-29.

Conneau, A., G. Lample, M. A. Ranzato, L. Denoyer and H. Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Conneau, A., G. Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H. and Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475-2485.

Conneau, A., A. Baevski, R. Collobert, A. Mohamed and M. Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Deshpande, A., T. Talukdar and K. Narasimhan. 2022. When is BERT multilingual? Isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3610–3623.

Dong, C., C. C. Loy, K. He and X. Tang. 2015. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 38*(2), 295-307.

Ellis, R. 2010. Second language acquisition, teacher education and language pedagogy. *Language Teaching 43(2),* 182-201.

Ettinger, A. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8, 24-48.

Futrell, R., E. Wilcox, T. Morita and R. Levy. 2019. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Giulianelli, M., J. Harding, F. Mohnert, D. Hupkes and W. Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 240–248.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1 (Long Papers),1195–1205.

Grimes, F, B. and J. E. Grimes. 2022. *Ethnologue*: *Languages of the World*. *Dallas, TX: SIL International.*

Hatch, E. M. 1983. *Psycholinguistics: A Second Language Perspective.* Rowley, MA: Newbury House Publishers, Inc.,

Hu, J., J. Gauthier, P. Qian, E. Wilcox and R. P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Huebner, P. A. and J. A. Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation,* 279-331.

Huebner, P. A., E. Sulem, F. Cynthia and D. Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning,* 624-646.

Kirov, C. and R. Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics* 6, 651-665.

Krashen, S. 1977. The monitor model for adult second language performance. *Viewpoints on English as a Second Language*, 152-161.

Krashen, S. 1981. Second language acquisition. *Second Language Learning* 3(7), 19-39.

Koo, K., J. Lee and M.-K. Park. 2022. An assessment of processing negative polarity Items by an L2 neural language model trained on English textbooks. *Language and Information Society* 46, 103-126.

Koo, K., J. Lee and M.-K. Park. 2023. Hierarchical inductive bias in the L2 textbook-T5 and Child-T5 language model: A study of data and architecture. *Korean Journal of Applied Linguistics* 39(4), 179-196.

Lakretz, Y., G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene and M. Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*Volume 1 (Long and Short Papers), 11–20.

Lample, G. and A. Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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

Manning, C. D. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics* 41(4), 701–707.

Niu, J. and G. Penn. 2020. Grammaticality and language modelling. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*.

Oba, M., T. Kuribayashi, H. Ouchi. and T. Watanabe. 2023. Second language acquisition of neural language models. *arXiv preprint arXiv:2306.02920*.

Papadimitriou, I. and D. Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models. *arXiv preprint arXiv:2004.14601*.

Pérez-Mayos, L., A. T. García, S. Mille and L. Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. *arXiv preprint arXiv:2105.04688*.

Pinker, S. and A. Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1-2), 73-193.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Ri, R. and Y. Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. *arXiv preprint arXiv:2203.10326*.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Rumelhart, D. E. and J. L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 216-271.

Sag, A. I., T. Wasow, M. E. Bender. 2003. *Syntactic Theory: A Formal Introduction*. Stanford: CSLI Publications.

Shi, F., X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, ... and D. Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 31210-31227.

Sportiche, D., H. Koopman and E. Stabler. 2013. *An Introduction to Syntactic Analysis and Theory*. John Wiley and Sons.

Sprouse, J. and N. Hornstein. 2013. Experimental syntax and island effects: Toward a comprehensive theory of islands. *Experimental Syntax and Island Effects*, 1-20.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Warstadt, A., A. Singh and S. R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics* 7, 625–641.

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.

Warstadt, A. and S. R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, 17-60.

Wilcox, E., R. Levy, T. Morita and R. Futrell. 2018. What do RNN language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Wu, S. and M. Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077*.

Wu, S., A. Conneau, H. Li, L. Zettlemoyer and V. Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *arXiv preprint arXiv:1911.01464*.

Yadavalli, A., A. Yadavalli and V. Tobin. 2023. SLABERT Talk pretty one day: Modeling second language acquisition with BERT. *arXiv preprint arXiv:2305.19589*.

Examples in: English
Applicable Languages: English
Applicable Level: All

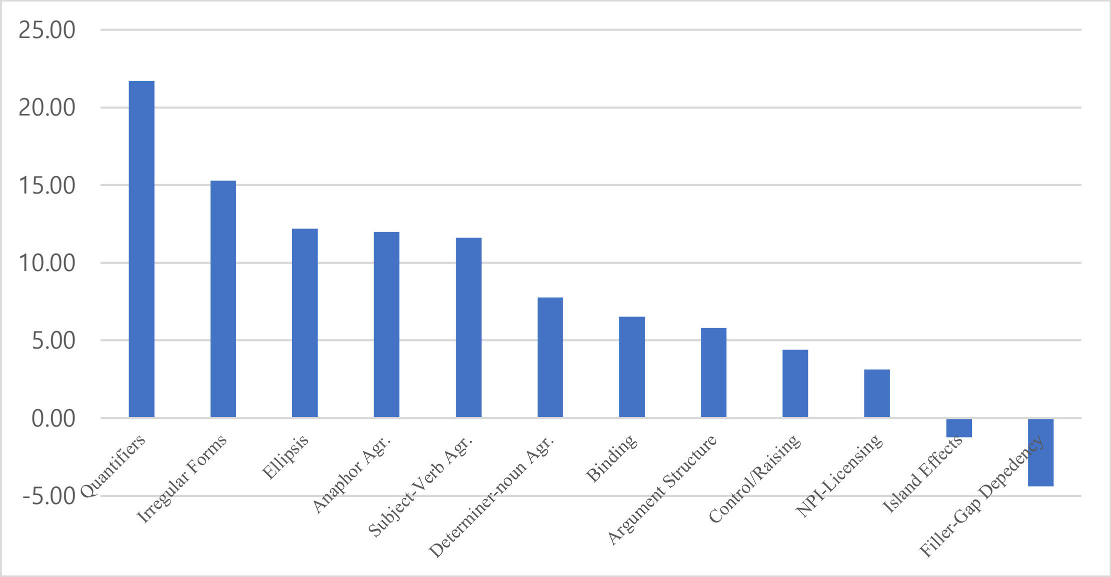
Appendices

Appendix 1. Performance of Model on All 67 Paradigms in the BLiMP

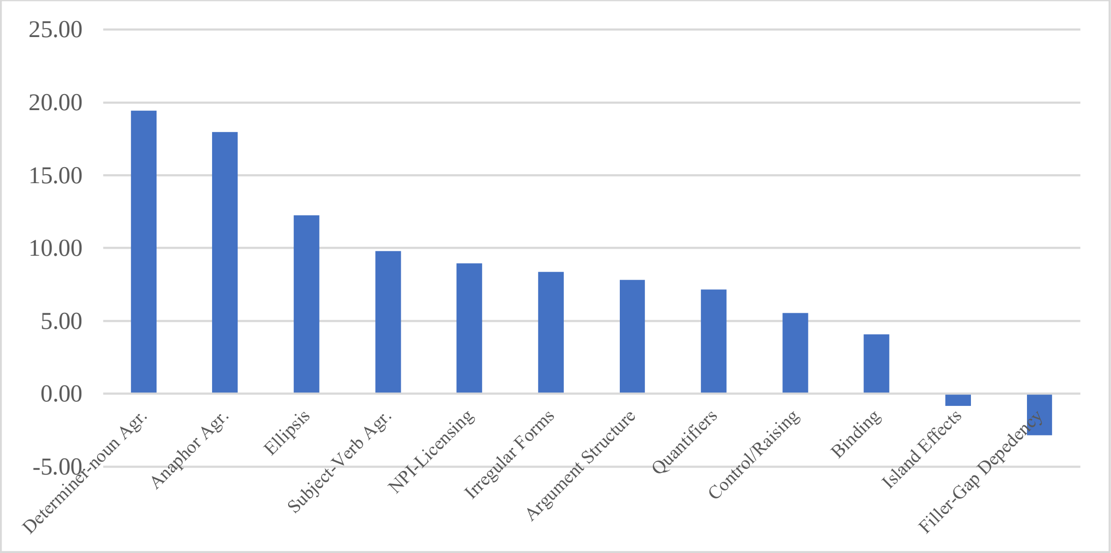
Phenomenom	Paradigm	En-Pre	En-L2	En-ADS	Ko-Pr	Ko-L2	Ko-ADS
anaphor agreement	anaphor_gender_agreement	82.3	58.59	61.25	30.8	54.79	47.74
	anaphor_number_agreement	93.6	81.9	75.35	40.2	61.72	52.93
argument structure	animate_subject_passive	59.4	64.73	61.33	53.3	60.23	65.48
	animate_subject_trans	80.1	72.88	73.12	28	45.13	59.31
	causative	63.7	65.72	66.62	37.7	51.45	45.78
	drop_argument	31.8	32.8	49.33	51.2	32.87	43.93
	inchoative	48	48.27	54.68	45.6	44.75	44.21
	intransitive	32.5	34.5	48.75	38.1	36.97	39.15
	passive_1	68.1	61.72	70.36	53.4	62.59	62.9
	passive_2	55.7	52.41	63.57	57.2	51.31	62.62
	transitive	72.8	67.54	69.55	55.6	62.88	63.46
Binding	principle_A_c_command	56.9	53.74	43.7	50.5	57.37	57.04
	principle_A_case_1	96	99.37	98.28	43.3	99.7	98.3
	principle_A_case_2	88.3	77.32	72.99	63.4	57.71	57.75
	principle_A_domain_1	97.9	88.91	73.48	0	90.13	75.14
	principle_A_domain_2	78.3	71.49	68.3	57.5	53.67	53.37
	principle_A_domain_3	57.5	54.78	54.91	47.9	50.36	49.82
	principle_A_reconstruction	32	38.81	47.75	14.5	29.88	39.47
control/rasing	existential_there_object_raising	74.5	63.67	60.7	65	64.64	60.34
	existential_there_subject_raising	71.7	67.98	60.48	60.2	53.48	46.7
	expletive_it_object_raising	68.7	65.38	60.83	59.6	61.79	60
	tough_vs_raising_1	34.2	44.87	40.22	42.2	41.99	49.97
	tough_vs_raising_2	76.8	65.71	75.7	58.8	63.77	53.21
determiner-noun agreement	determiner_noun_agreement_1	72.4	72.74	83.37	49.4	67.54	64.99
	determiner_noun_agreement_2	82.7	77.67	81.42	52.5	68.26	62.44
	determiner_noun_agreement_irregular_1	60.6	60.45	70.8	48.4	61.82	57.25
	determiner_noun_agreement_irregular_2	79.2	70.91	80.35	55.9	68.24	66.77
	determiner_noun_agreement_with_adjective_1	72.7	70.46	76.6	49.1	57.42	53.35
	determiner_noun_agreement_with_adj_2	65.3	65.58	77.97	43.5	57.22	54.13
	determiner_noun_agreement_with_adj_irregular_1	70.9	71.14	75.51	53.7	58.77	54.74
	determiner_noun_agreement_with_adj_irregular_2	65.6	70.83	75.59	52.5	58.46	52.56
ellipsis	ellipsis_n_bar_1	56.2	56.83	51.1	41.6	46.69	32.08
	ellipsis_n_bar_2	66.2	63.57	75.36	88.1	49.31	69.88
filler gap	wh_questions_object_gap	64.4	56.33	47.34	11.8	63.22	55.5
	wh_questions_subject_gap	85.6	66.17	70.01	34.2	76.86	83.07
	wh_questions_subject_gap_long_distance	82.9	68.48	69.27	42.6	87.22	85.21
	wh_vs_that_no_gap	95.6	88.28	72.18	35.1	89.77	82.05
	wh_vs_that_no_gap_long_distance	97.1	90.29	79.82	33.5	87.3	82.99
	wh_vs_that_with_gap	7.9	16.66	40.02	56.5	10.14	23.15
	wh_vs_that_with_gap_long_distance	5	10.99	33.37	56.6	13.34	19.85
irregular forms	irregular_past_participle_adjectives	97.6	75.8	71.64	95.8	61.69	57.46
	irregular_past_participle_verbs	92.9	89.74	70.26	23.5	73.32	67.72
island effects	adjunct_island	30.5	35.25	69.28	0	48.52	65.29
	complex_NP_island	43.5	44.86	55.14	44.9	45.49	51.39
	coordinate_structure_constraint_complex_left_branc	34.3	41.57	31.5	22.5	33.91	32.13
	coordinate_structure_constraint_object_extraction	45.4	43.02	55.32	53	46.09	46.05
	left_branch_island_echo_question	87.5	77.02	44.45	98.4	76.66	62.01
	left_branch_island_simple_question	53.4	62.24	35.73	30.8	59.91	37.26
	sentential_subject_island	63.1	55.05	46.02	34.4	46.2	56.51
	wh_island	27.3	30.27	55.28	0.1	42.25	48.85
npi licensing	matrix_question_npi_licensor_present	3.2	1.2	43.57	8	2.44	20.4
	npi_present_1	67.2	65.8	50.08	96.4	53.49	53.96
	npi_present_2	72.3	69.15	59.73	95.1	57.05	55.22
	only_npi_licensor_present	66.3	55.09	68.23	86.2	36.14	41.05
	only_npi_scope	54.1	45.29	60.64	40.3	42.48	57.83
	sentential_negation_npi_licensor_present	98.5	91.34	95.33	60.6	99.36	88.48
	sentential_negation_npi_scope	28.9	23.84	51.43	58.3	38.79	49.45
quantifiers	existential_there_quantifiers_1	93.3	96.94	86.51	30	75.56	70.48
	existential_there_quantifiers_2	79.9	77.97	43.97	13.9	46.08	53.16
	superlative_quantifiers_1	79.1	88.57	85.8	65.1	82.88	92.55
	superlative_quantifiers_2	67.1	87.66	79.36	0	59.81	50.85
subject verb agreement	distractor_agreement_relational_noun	72	68.88	58.71	43.2	44.91	46.76
	distractor_agreement_relative_clause	60.6	52.46	49.35	45.9	43.63	42.62
	irregular_plural_subject_verb_agreement_1	68.9	65.04	61.18	53.9	57.26	52.39
	irregular_plural_subject_verb_agreement_2	73.1	67.48	65.31	49.9	63.01	55.67
	regular_plural_subject_verb_agreement_1	72.9	68.21	64.32	48.6	54.1	54.7
	regular_plural_subject_verb_agreement_2	74	67.88	64.56	51.8	57.39	52.43
Average		65.04477612	62.06104	63.19447761	46.41194	56.10716	55.98955224

Appendix 2. Performance of Model on All 67 Paradigms of Four Category

Morphology	Paradigm	En-Pre	En-L2	En-ADS	Ko-Pre	Ko-L2	Ko-ADS
anaphor agreement	anaphor_gender_agreement	82.3	58.59	61.25	30.8	54.79	47.74
	anaphor_number_agreement	93.6	81.9	75.35	40.2	61.72	52.93
determiner-noun agreement	determiner_noun_agreement_1	72.4	72.74	83.37	49.4	67.54	64.99
	determiner_noun_agreement_2	82.7	77.67	81.42	52.5	68.26	62.44
	determiner_noun_agreement_irregular_1	60.6	60.45	70.8	48.4	61.82	57.25
	determiner_noun_agreement_irregular_2	79.2	70.91	80.35	55.9	68.24	66.77
	determiner_noun_agreement_with_adjective_1	72.7	70.46	76.6	49.1	57.42	53.35
	determiner_noun_agreement_with_adj_2	65.3	65.58	77.97	43.5	57.22	54.13
	determiner_noun_agreement_with_adj_irregular_1	70.9	71.14	75.51	53.7	58.77	54.74
	determiner_noun_agreement_with_adj_irregular_2	65.6	70.83	75.59	52.5	58.46	52.56
irregular forms	irregular_past_participle_adjectives	97.6	75.8	71.64	95.8	61.69	57.46
	irregular_past_participle_verbs	92.9	89.74	70.26	23.5	73.32	67.72
subject verb agreement	distractor_agreement_relational_noun	72	68.88	58.71	43.2	44.91	46.76
	distractor_agreement_relative_clause	60.6	52.46	49.35	45.9	43.63	42.62
	irregular_plural_subject_verb_agreement_1	68.9	65.04	61.18	53.9	57.26	52.39
	irregular_plural_subject_verb_agreement_2	73.1	67.48	65.31	49.9	63.01	55.67
	regular_plural_subject_verb_agreement_1	72.9	68.21	64.32	48.6	54.1	54.7
	regular_plural_subject_verb_agreement_2	74	67.88	64.56	51.8	57.39	52.43
	Average	75.40556	69.7644	70.19667	49.36667	59.41944	55.36944
Syntax							
argument structure	animate_subject_passive	59.4	64.73	61.33	53.3	60.23	65.48
	animate_subject_trans	80.1	72.88	73.12	28	45.13	59.31
	causative	63.7	65.72	66.62	37.7	51.45	45.78
	drop_argument	31.8	32.8	49.33	51.2	32.87	43.93
	inchoative	48	48.27	54.68	45.6	44.75	44.21
	intransitive	32.5	34.5	48.75	38.1	36.97	39.15
	passive_1	68.1	61.72	70.36	53.4	62.59	62.9
	passive_2	55.7	52.41	63.57	57.2	51.31	62.62
	transitive	72.8	67.54	69.55	55.6	62.88	63.46
ellipsis	ellipsis_n_bar_1	56.2	56.83	51.1	41.6	46.69	32.08
	ellipsis_n_bar_2	66.2	63.57	75.36	88.1	49.31	69.88
filler gap	wh_questions_object_gap	64.4	56.33	47.34	11.8	63.22	55.5
	wh_questions_subject_gap	85.6	66.17	70.01	34.2	76.86	83.07
	wh_questions_subject_gap_long_distance	82.9	68.48	69.27	42.6	87.22	85.21
	wh_vs_that_no_gap	95.6	88.28	72.18	35.1	89.77	82.05
	wh_vs_that_no_gap_long_distance	97.1	90.29	79.82	33.5	87.3	82.99
	wh_vs_that_with_gap	7.9	16.66	40.02	56.5	10.14	23.15
	wh_vs_that_with_gap_long_distance	5	10.99	33.37	56.6	13.34	19.85
island effects	adjunct_island	30.5	35.25	69.28	0	48.52	65.29
	complex_NP_island	43.5	44.86	55.14	44.9	45.49	51.39
	coordinate_structure_constraint_complex_left_branch	34.3	41.57	31.5	22.5	33.91	32.13
	coordinate_structure_constraint_object_extraction	45.4	43.02	55.32	53	46.09	46.05
	left_branch_island_echo_question	87.5	77.02	44.45	98.4	76.66	62.01
	left_branch_island_simple_question	53.4	62.24	35.73	30.8	59.91	37.26
	sentential_subject_island	63.1	55.05	46.02	34.4	46.2	56.51
	wh_island	27.3	30.27	55.28	0.1	42.25	48.85
	Average	56.07692	54.1327	57.25	42.46923	52.73308	54.61962
Semantics							
npi licensing	matrix_question_npi_licensor_present	3.2	1.2	43.57	8	2.44	20.4
	npi_present_1	67.2	65.8	50.08	96.4	53.49	53.96
	npi_present_2	72.3	69.15	59.73	95.1	57.05	55.22
	only_npi_licensor_present	66.3	55.09	68.23	86.2	36.14	41.05
	only_npi_scope	54.1	45.29	60.64	40.3	42.48	57.83
	sentential_negation_npi_licensor_present	98.5	91.34	95.33	60.6	99.36	88.48
	sentential_negation_npi_scope	28.9	23.84	51.43	58.3	38.79	49.45
quantifiers	existential_there_quantifiers_1	93.3	96.94	86.51	30	75.56	70.48
	existential_there_quantifiers_2	79.9	77.97	43.97	13.9	46.08	53.16
	superlative_quantifiers_1	79.1	88.57	85.8	65.1	82.88	92.55
	superlative_quantifiers_2	67.1	87.66	79.36	0	59.81	50.85
	Average	64.53636	63.8955	65.87727	50.35455	54.00727	57.58455
Syntax-Semantic interface							
Binding	principle_A_c_command	56.9	53.74	43.7	50.5	57.37	57.04
	principle_A_case_1	96	99.37	98.28	43.3	99.7	98.3
	principle_A_case_2	88.3	77.32	72.99	63.4	57.71	57.75
	principle_A_domain_1	97.9	88.91	73.48	0	90.13	75.14
	principle_A_domain_2	78.3	71.49	68.3	57.5	53.67	53.37
	principle_A_domain_3	57.5	54.78	54.91	47.9	50.36	49.82
	principle_A_reconstruction	32	38.81	47.75	14.5	29.88	39.47
control/rasing	existential_there_object_raising	74.5	63.67	60.7	65	64.64	60.34
	existential_there_subject_raising	71.7	67.98	60.48	60.2	53.48	46.7
	expletive_it_object_raising	68.7	65.38	60.83	59.6	61.79	60
	tough_vs_raising_1	34.2	44.87	40.22	42.2	41.99	49.97
	tough_vs_raising_2	76.8	65.71	75.7	58.8	63.77	53.21
	Average	69.4	66.0025	63.11167	46.90833	60.37417	58.42583



Appendix 3. Performance Drop between En-Textbook and Ko-Textbook Models by the 12 Phenomena



Appendix 4. Performance Drop between En-ADS and Ko-ADS Models by the 12 Phenomena