Evaluating L2 Training Methods in Neural Language Models

Jaemin Lee & Jeong-Ah Shin[†]
Dongguk University

ABSTRACT
Recent advancements in language models (LMs) have significantly improved language processing capabilities; however, these models remain less efficient than human learning, especially when trained on developmentally plausible data volumes similar to those encountered by children (Warstadt & Bowman, 2022; Linzen, 2020). The inefficiency is even more pronounced in second language (L2) acquisition contexts, where cross-linguistic transfer is a key phenomenon (Papadimitriou & Jurafsky, 2020; Yadavalli et al., 2023). This study evaluates L2 training methods in neural language models by examining mutual L1-L2 influences during learning with developmentally plausible data volumes. We propose two approaches to mitigate catastrophic forgetting: the One-Stage Training (OST) method, which integrates L1 and L2 learning into a single stage, and the One-Stage Mixed Training (OSMT) method, which refines OST by incorporating L1 data into the L2 stage for more realistic simulation of bilingual learning. Through continuous syntactic evaluations throughout training, we analyzed how L1 performance changes during L2 acquisition and how cross-linguistics transfer emerges in Korean and English. The results indicate that OST and OSMT effectively mitigated catastrophic forgetting and supported more stable learning compared to the conventional Two-Stage Training method. OSMT achieved superior integration of L1 and L2 structures while revealing negative transfer effects from Korean (L1) to English (L2). These findings provide valuable insights into both neural model training and human-like L2 acquisition processes.

Keywords: developmentally plausible data, cross-linguistic transfer, second language acquisition, neural language models, L2 language models, catastrophic forgetting

1. Introduction

For decades, language models have been used to simulate human language learning and processing (Elman, 1990; Hale, 2001; Reali & Christiansen, 2005). With advances in deep learning, artificial neural-network language models have been advanced by scaling up both training data and model parameters. Models such as

[†] Corresponding author: jashin@dongguk.edu

https://doi.org/10.30961/lr.2024.60.3.323

GPT-3 (Brown et al., 2020) exemplify this trend, using enormous datasets and extensive computational power to achieve high performance across various tasks. However, Warstadt and Bowman (2022) and Linzen (2020) argue that the sheer volume of data is the primary advantage modern language models have over humans. When confined to developmentally plausible data volumes, these models significantly underperform on benchmarks evaluating human-like syntactic and semantic behavior (van Schijndel, 2019; Zhang et al., 2020). Recognizing these limitations has led researchers to explore alternative approaches that mirror the efficiency of human language acquisition.

Building on the idea of using developmentally plausible data volumes for training, Huebner et al. (2021) introduced BabyBERTa, a scaled-down version of the RoBERTa model, designed to simulate the language input available to children between the ages of one and six. BabyBERTa was trained on the AO-CHILDES corpus, which consists of approximately five million words of American-English transcribed child-directed speech. The model employs dynamic masking and hyper-parameters optimized for small-scale language acquisition experiments, using only eight layers, eight attention heads, 256 hidden units, and an intermediate size of 1,024. By duplicating input sequences and applying novel random masks, BabyBERTa effectively receives a more extensive language experience, analogous to that of a six-year-old child. Despite having fifteen times fewer parameters and 6,000 times fewer words than RoBERTa-base, BabyBERTa demonstrated comparable grammatical knowledge acquisition. The results highlighted that the model could achieve high performance with significantly less data, underscoring the potential for using child-directed language to develop efficient language models.

Following the success of BabyBERTa, the BabyLM Challenge (Warstadt et al., 2023) was introduced to further investigate and promote the development of language models trained on child-directed data. The challenge encourages the creation of models that can achieve high performance with limited and contextually rich data, much like the data available to human children. By focusing on smaller, more specialized datasets and leveraging insights from cognitive science, the movement to build developmentally plausible models not only pave the way for more efficient and potentially more robust language models, but also offer profound insights into human language acquisition.

Related to human language acquisition and development, an interesting topic that recent studies of natural language processing (NLP) have paid attention to is cross-linguistic transfer effects, where a speaker's first language (L1) influences second language (L2) acquisition (e.g., Conneau et al., 2018). Cross-linguistic transfer, a key concept in linguistics and cognitive science, occurs as structural or lexical features from L1 affect learning in L2. This transfer can be positive, aiding learning when languages share structures, or negative, creating challenges when there are significant linguistic differences (Jarvis & Pavlenko, 2007). A long-standing debate on negative transfer centers on the Representational Deficit Approach and the Computational Difficulty Approach. The Representational Deficit Approach links poor L2 performance to incomplete or permanently deficient L2 knowledge, evidenced by advanced learners' struggles with morphology and grammaticality judgments (Hawkins & Chan, 1997; Johnson & Newport, 1989, 1991). Conversely, the Computational Difficulty Approach attributes errors to performance issues rather than impaired syntax, aligning with the Missing Surface Inflection Hypothesis (Haznedar & Schwartz, 1997; Prévost & White, 2000). This view emphasizes the role of L1 in performance challenges, suggesting that learners' errors reflect processing difficulties rather than fundamental knowledge deficits.

This cross-linguistic transfer has been expanded to research on neural models. For instance, Papadimitriou and Jurafsky (2020) found that inductive biases acquired from diverse linguistic datasets as well as non-linguistic datasets with distinct patterns, such as music and code, can enhance language model learning, and Yadavalli et al. (2023) showed that conversational data facilitated language acquisition more than scripted data, with negative transfer increasing with linguistic distance. Oba et al. (2023) observed that pretraining on L1 improves L2 learning, especially for syntax and morphology, depending on language similarity. Research also explored Korean-English syntactic transfer (Koo et al., 2024), adding insights into cross-linguistic challenges with these structurally diverse languages.

These studies employ various methods for building L2 models, typically classified into fine-tuning and continual learning. Papadimitriou and Jurafsky (2020) and Yadavalli et al. (2023) fine-tuned the model after L1 learning for specific purposes. In other words, these studies aimed to adjust only the word embeddings for L2 while maintaining the inductive bias of L1. In contrast, continual learning, as seen in Oba et al. (2023) and Koo et al. (2024), involves sequential L1 and L2 learning stages. Both approaches adopt a Two-Stage Training process where L1 learning is completed before L2 begins, reusing the model across stages.

However, these methods expose L2 models to catastrophic forgetting, where L1 knowledge is overwritten during L2 learning (Li & Hoiem, 2017; Lopez-Paz & Ranzato, 2017). Catastrophic forgetting occurs when neural network representations for new tasks interfere with those formed for prior tasks due to shared neural resources (Kemker et al., 2018; Kaushik et al., 2021). This stability-plasticity dilemma, where the model must balance retaining old knowledge and learning new information, poses a challenge in effective L2 learning without memory loss of L1 (Kirkpatrick et al., 2016).

The impact of catastrophic forgetting questions the validity of the L2 model itself. Natural language acquisition suggests that L2 is built on a foundation of L1 knowledge (McManus, 2021). If L1 is largely forgotten and only recent L2 knowledge is retained, it challenges the core concept of L2 acquisition as a continuation of L1-based learning. Furthermore, while models may retain some L1 inductive bias, this does not guarantee a true knowledge transfer to L2, making catastrophic forgetting a critical issue in designing L2 models. Therefore, to develop more plausible L2 models and get a better understanding of cross-linguistic transfer, it is essential to address catastrophic forgetting by considering L2 training methods that preserve both L1 and L2 knowledge effectively.

This study aims to evaluate various training methods for L2 learning within neural language models, specifically focusing on mitigating catastrophic forgetting, a phenomenon where previous L1 knowledge degrades during L2 learning. Inspired by the ‘scaffolding’ experiment of Huebner et al. (2021), which showed that the ordering of training data in a single training stage could significantly influence grammatical knowledge retention and development even when the model ultimately learns the exact same data, suggesting that starting with easier data could serve as a better foundation for learning more complex data, this study proposes two novel methods: One-Stage Training (OST) and One-Stage Mixed Training (OSMT). OST involves learning L1 followed immediately by L2 within a single training stage, while OSMT builds on OST by interweaving L2 training data with L1, facilitating blended dual-language exposure. By observing models trained using these methods alongside the conventional TST approach, the study tracks shift in syntactic knowledge and investigates cross-linguistic transfer effects between Korean and English. The research questions are as follows:

Q1. How does L1 performance change as L2 learning progresses under different training methods, and what does this reveal about the phenomenon of catastrophic forgetting?

Q2. How does L2 performance develop during these training approaches, and in what ways does it relate to catastrophic forgetting?

Q3. How does cross-linguistic transfer between Korean and English emerge in shared and distinct syntactic paradigms across both languages?

We hypothesize that catastrophic forgetting will appear in the conventional TST method with rapid L1 performance decline, while OST and OSMT should show gradual declines as Huebner et al.’s (2021) scaffolding experiment demonstrated that the data learned earlier influences the model's ability to process data learned later and starting with easier data could serve as a better foundation for learning more complex data. Furthermore, considering Yadavalli et al. (2023), which demonstrated that greater linguistic distance leads to stronger negative transfer effects, along with the significant linguistic distance between Korean and English (Chiswick & Miller, 2005), the conventional Two-Stage Training is expected to show faster L2 performance gains, whereas OST and OSMT should demonstrate more gradual improvements, suggesting stronger L1 influence. The linguistic distance between Korean and English is expected to negatively impact L2 learning, leading to negative transfer effects in most of the paradigms, although positive transfer effects are also expected in some paradigms where both languages are syntactically similar.

2. Experiments

We conducted two experiments to evaluate the impact of different L2 training methods on cross-linguistic transfer and catastrophic forgetting in neural language models. Experiment 1 set L1 as English and L2 as Korean, focusing on L1 syntactic retention as L2 learning progresses. Experiment 2 set L1 as Korean and L2 as English, evaluating how prior L1 knowledge influences L2 learning. An additional model trained solely on English was included for baseline comparison, tracking L2 English performance changes as L2 training progressed.

2.1. Methods

This study tested three L2 training methods to examine L1 retention and patterns of L2 acquisition (see Figure 1):

a. Conventional Two-Stage Training (TST): The model first learns L1, followed by L2 in separate stages. In this model, once L1 learning is fully completed and the initial training process has terminated, the trained L1 model is loaded as the starting point for L2 learning. Therefore, the L1 model serves as the initial state when L2 learning begins. Common in prior L2 research, it often faces catastrophic forgetting, where L1 knowledge is overwritten during L2 learning.

b. One-Stage Training (OST): L1 and L2 are learned sequentially within a single stage. In OST, training does not end upon the completion of L1 learning; instead, the data is switched to L2, and training continues seamlessly. This approach aims to avoid catastrophic forgetting by training both languages continuously, facilitating greater cross-linguistic transfer.

c. One-Stage Mixed Training (OSMT): A variation of OST, in which L2 learning includes mixed L1 and L2 inputs. This simulates real-world L2 acquisition more closely by reinforcing L1 knowledge during L2 learning. For example, with a 5-million-word dataset for L1 and 1 million for L2, OSMT merges the last portion of L1 data with L2 data, resulting in a 6-million-word dataset.

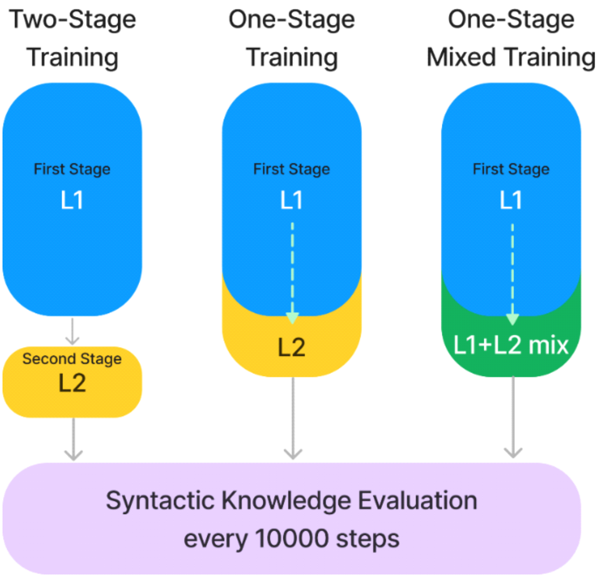

Figure 1. Overall structure of two experiments

Using these methods, Experiment 1 focused on L1 English syntactic retention as L2 Korean learning progresses. The syntactic performance of L1 was measured every

10,000 training steps to track L1 retention during L2 acquisition. Experiment 2 evaluated how prior L1 Korean knowledge influences L2 English learning. An additional model which was trained solely and normally on English was included for baseline comparison[1].

2.2. Model Training and Evaluation for Second Language Acquisition

2.2.1. Model

The BabyBERTa model (Huebner et al., 2021), a scaled-down version of RoBERTa, was chosen for its effectiveness in syntactic learning with limited data. BabyBERTa uses dynamic masking to create different masking patterns for repeated sentences, simulating human-like exposure with a small dataset of 5 million words. This dynamic masking prevents simple data repetition, making BabyBERTa suitable for sequential and continuous L1 and L2 learning in our experiments. Given its alignment with human-like language acquisition, BabyBERTa is well-suited for studying L1-L2 transfer effects.

The original BabyBERTa model utilized a monolingual vocabulary of 8,912 words. To create a bilingual tokenizer, we applied Byte-Pair Encoding (BPE) to both L1 and L2 data, expanding the vocabulary size to 16,384 words, effectively doubling its capacity to accommodate both languages. In TST, an L1-specific tokenizer was used initially, then replaced with a bilingual tokenizer for L2 training. OST and OSMT employed the bilingual tokenizer throughout all training stages.

2.2.2. Data

For all experiments, English was the standard evaluation language. In Experiment 1, L1 was set as English and L2 as Korean, with English data from the AO-CHILDES corpus (5 million words) and Korean data from K-CDS (1.2 million words). Experiment 2 used Korean as L1 and English as L2, with Korean data from the Modu Corpus (4.5 million words) from National Institute of Korean Language (2020) and English data from L2-textbook (1.5 million words). In each experiment,

[1] The code to reproduce our experiments and the data used for training and evaluation can be found at https://github.com/jeongahshin/babyLM_L2.

L1 data constituted around 75-80% of the training data, simulating typical second language acquisition where L1 is predominant. The overall size of the datasets used in the two experiments is similar. However, there is a significant difference in the total training steps: 230k for Experiment 1 and 150k for Experiment 2. This discrepancy arises because, although the datasets are comparable in terms of the number of words, the length of sentences comprising the datasets differs. In other words, since the model typically processes one sentence as a unit of training, even though the number of words is the same, the sentences in the dataset used for Experiment 2 are longer than those in Experiment 1.

2.2.3 Syntactic Knowledge Evaluation

The Zorro test set (Huebner et al., 2021), an adaptation of the BLiMP benchmark (Warstadt et al., 2020), was used to evaluate syntactic knowledge. Zorro assesses specific syntactic phenomena with minimal pairs of grammatical and ungrammatical sentences, each pair designed to test a distinct syntactic structure (e.g., subject-verb agreement). The model scores sentence pairs based on cross-entropy error, with accuracy calculated as the proportion of pairs correctly judged. In contrast to BLiMP, which imposed no specific limitations on vocabulary, Huebner et al. (2021) constructs test sentence pairs using the restricted vocabulary of BabyBERTa. This streamlined vocabulary ensures that the model's syntactic performance is not undervalued due to constraints in its lexical capabilities.

Zorro encompasses 13 syntactic phenomena and 23 paradigms, with each paradigm containing 4,000 sentences (2,000 pairs). This syntactic evaluation allows for consistent tracking of L1 and L2 knowledge retention and cross-linguistic transfer effects throughout the experiments.

3. Results

3.1. Experiment 1: L1 Retention under L2 Training

3.1.1. General Results and Analysis

Experiment 1 examined the effects of different L2 training methods—Two-Stage

Training (TST), One-Stage Training (OST), and One-Stage Mixed Training (OSMT)—on the retention of L1 knowledge and acquisition of L2 syntactic competence. It evaluated how each method influences cross-linguistic transfer and catastrophic forgetting, with the focus on syntactic knowledge transfer between English (L1) and Korean (L2).

Figure 2 shows the learning curves for English (L1) syntactic performance across three phases, represented by different training steps. Initially, up to the 150k step, all methods trained exclusively on English, with L1 accuracy increasing similarly across TST, OST, and OSMT (TST: 0.705, OST: 0.712, OSMT: 0.712). During the intermediate phase (150k-190k), OSMT began L2 Korean training while continuing with L1 English, resulting in a learning curve similar to TST and OST, which remained focused on English only (TST: 0.744, OST: 0.751, OSMT: 0.749).

In the final phase (post-190k steps), TST and OST transitioned to L2 Korean, showing significant divergence. TST exhibited a sharp drop in L1 English accuracy, reflecting catastrophic forgetting, while OST's performance declined more gradually. In contrast, OSMT maintained stable L1 performance, achieving the highest accuracy among the methods by the end of training (TST: 0.510, OST: 0.631, OSMT: 0.754).

These trends suggest that OSMT's continuous exposure to both languages helps preserve L1 syntactic knowledge, while TST's separate training stages hinder L1 retention due to abrupt learning shifts.

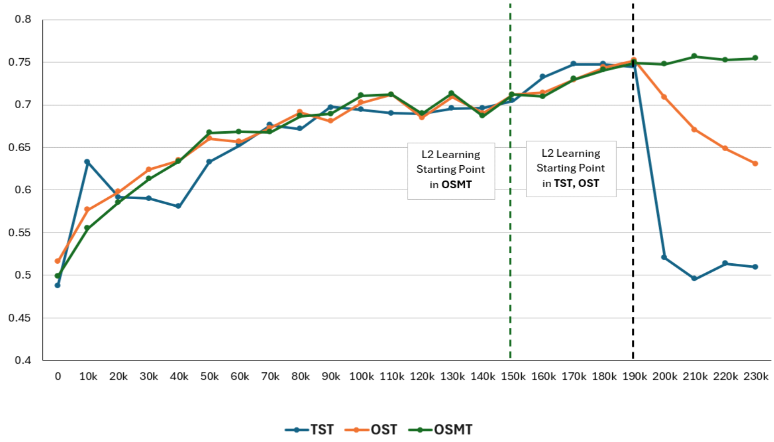

Figure 2. Learning curve of L1 English

3.1.2. Paradigm-Specific Analysis

To analyze specific syntactic patterns, Figure 3 shows accuracy by paradigm. While general learning curves across paradigms resemble the trends in Figure 2, certain paradigms show unique patterns; for example, OST outperforms OSMT in agreement_subject_verb, while OSMT excels in superlative quantifiers and binding_principle_a. These variations suggest that mixed L2 exposure affects specific L1 syntactic structures differently.

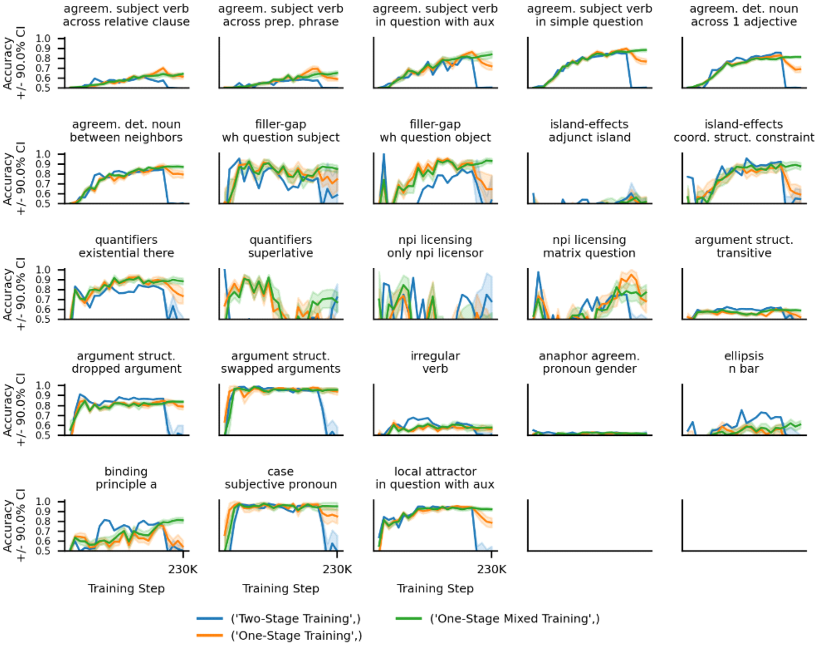

Figure 3. Learning curve of L1 English by paradigm

To statistically test these effects, critical points at 190k (L1 training ending point in TST) and 230k training steps were established. These points serve to highlight specific stages of language learning, allowing for a clearer comparison of L1 performance under distinct training conditions. To verify the reliability of these calculated accuracy differences, paired t-tests were conducted. The significance level was set at 0.05, and items with a p-value lower than this threshold (*p*<0.05) were marked with an asterisk.

At the 190k step, the OST model had completed L1 English training, representing a pure L1 English model, while the OSMT model was simultaneously learning L1 English and L2 Korean. This allows a direct comparison to observe how mixed L1-L2 exposure impacts syntactic retention in contrast to an L1-exclusive learning stage. By 230k steps, OSMT had accumulated more total training data, yet the amount of English data remained constant from 190k, controlling for the effect of additional L2 exposure on L1 performance without an increase in L1 data.

In the comparison with the OSMT model at the 190k step, significant accuracy decreases, interpreted as negative transfer effects, were observed in six paradigms, including all agreement_subject_verb and agreement_determiner_noun paradigms (e.g., across_1_adjective) and argument_structure (e.g., dropped_argument). No significant increases in accuracy, or positive transfer effects, were observed. This suggests that L2 Korean may influence the formation of L1 English syntactic knowledge. However, it may also reflect reduced exposure to L1 English.

Further analyses comparing the OSMT model at 190k and 230k steps with OST at 190k allowed us to control for L1 English training data scarcity effects. As shown in Table 1, at 230k, significant negative transfer effects were limited to two paradigms in OSMT (agreement_subject_verb: across_relative_clause, across_prepositional_phrase), and a single significant positive effect was observed in binding_principle_a. While the overall trends of performance increases and decreases remained similar, the magnitude of decreases lessened, with some differences becoming statistically insignificant.

Certain paradigms, such as superlative in quantifiers and all npi_licensing paradigms, exhibited substantial performance changes, though these differences were not statistically significant. As shown in Figure 3, these paradigms displayed fluctuating learning curves, contrasting with the stable trends seen in other paradigms. This indicates that the model struggles to generalize syntactic knowledge in these specific paradigms, potentially due to their complexity or limited representational consistency in the training data.

Table 1. Performance difference between critical points

Phenomena	Paradigms	OSMT(190k) vs. OST(190k)	OSMT(230k) vs. OST(190k)
agreement_ subject_verb	across_relative_clause	-0.059*	-0.058*
	across_prepositional_phrase	-0.052*	-0.042*
	in_question_with_aux	-0.061*	-0.025
	in_simple_question	-0.037*	-0.014

Table 1. Continued

Phenomena	Paradigms	OSMT(190k) vs. OST(190k)	OSMT(230k) vs. OST(190k)
agreement_ determiner_ noun	across_1_adjective	-0.024*	-0.016
	between_neighbors	-0.009	-0.007
filler-gap	wh_question_subject	-0.013	-0.004
	wh_question_object	-0.020	0.020
island-effects	adjunct_island	0.023	0.034
	coordinate_structure_constraint	0.022	-0.001
quantifiers	existential_there	0.001	-0.003
	superlative	0.088	0.116
npi_licensing	only_npi_licensor	0.198	0.124
	matrix_question	-0.095	-0.110
argument_ structure	transitive	0.009	0.002
	dropped_argument	-0.031*	-0.010
	swapped_arguments	-0.020	-0.020
irregular	verb	0.007	0.003
anaphor_ agreement	pronoun_gender	0.004	-0.003
ellipsis	n_bar	0.008	0.049
binding	principle_a	0.040	0.059*
case	subjective_pronoun	-0.031	-0.020

*p<.05

3.2 Experiment 2

3.2.1 General Results and Analysis

Experiment 2 shifts focus, with Korean as L1 and English as L2, to analyze L1 impact on L2 acquisition. Figure 4 presents L2 English learning curves across three phases. Up to 90k steps, all models focus on Korean, yielding no significant L2 gains. Starting at 90k, OSMT introduces L2 English, showing gradual improvements from 0.529 to 0.614. The other methods transition to L2 English after 120k steps, with all models showing enhanced performance: TST exhibits the steepest gain, OST shows moderate improvement, and OSMT reaches the highest final accuracy (0.663).

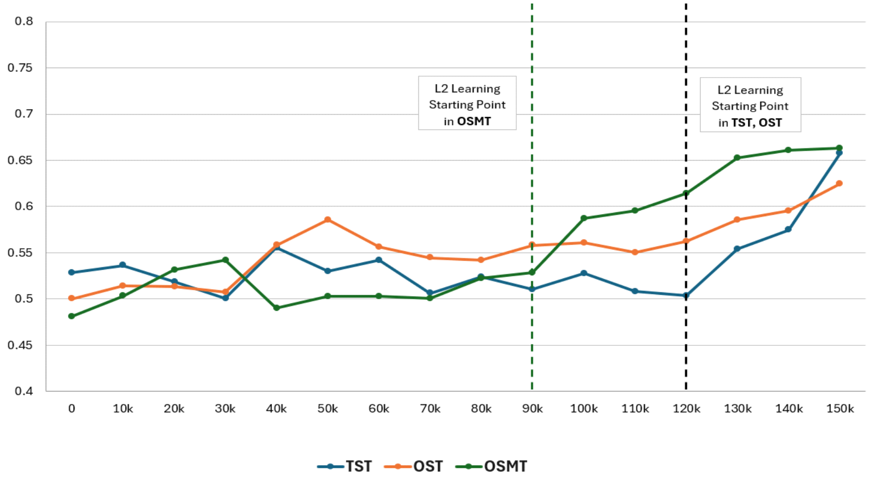

Figure 4. Learning curve of L2 English

Comparison with an English-only model (En(L1)) reveals that TST's learning curve closely resembles that of a monolingual learner (see Figure 5), suggesting TST's limited L1 retention has little impact on L2 syntactic stability. OST and OSMT's gradual improvement suggest they maintain L1 influence, with OSMT's mixed exposure supporting a balanced learning curve across both languages.

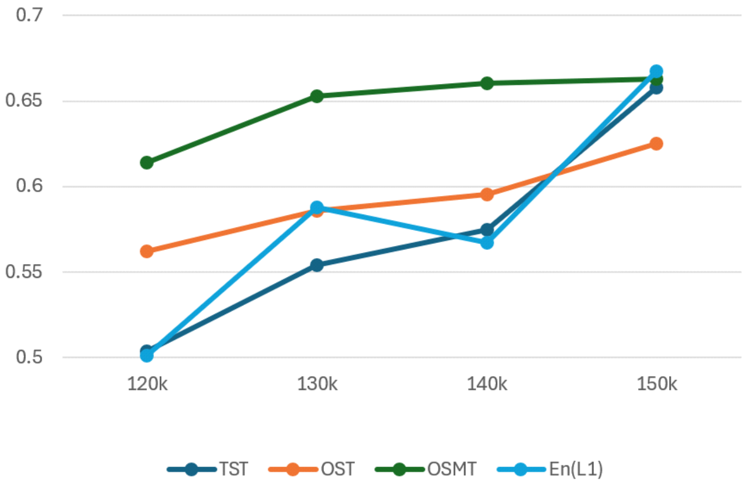

Figure 5. Final section with en(L1) English model

3.2.2 Paradigm-Specific Analysis

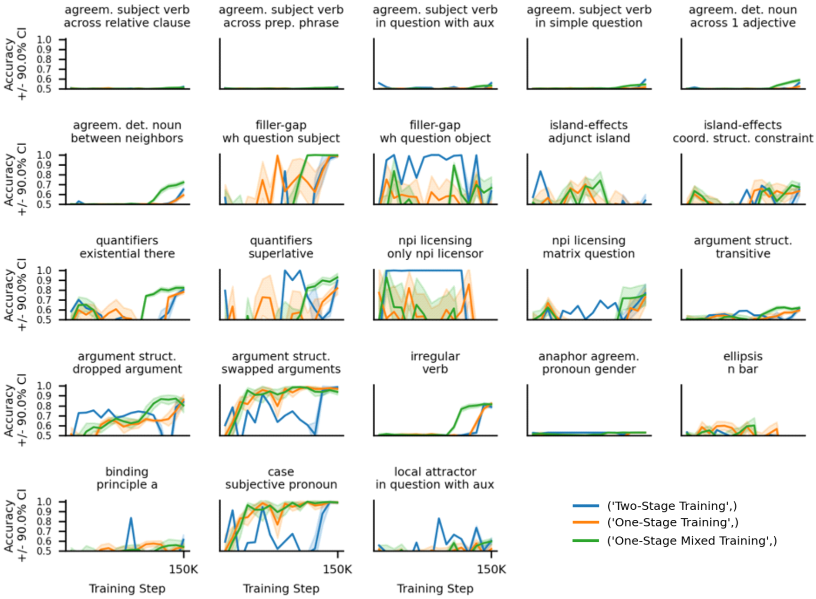

Figure 6. Learning curve of L2 English by paradigm

Figure 6 presents paradigm-specific accuracy for L2 English acquisition, revealing complex patterns not observed in Experiment 1. In contrast to the generally stable trends across paradigms in Figure 3 in Experiment 1, Figure 6 shows sharp fluctuations in accuracy. When limiting the analysis to post-120k training steps, specific paradigms—such as agreement_determiner_noun, quantifiers, and the transitive paradigm within argument_structure—align more closely with the average trend, indicating gradual improvement in syntactic performance under different training methods.

To investigate the transfer effects more rigorously, we compared the performance of TST, OST, and OSMT with a monolingual English model (En(L1)) as a baseline, allowing us to control for L1 Korean's influence on L2 English learning (see Table 2). The results indicate significant negative transfer effects in paradigms such as agreement_subject_verb and argument_structure, where Korean's syntactic structures diverge from English. OST exhibited negative transfer in 11 paradigms, reflecting the influence of Korean syntactic interference during L2 English learning. In contrast, TST reduced negative transfer effects to only five paradigms, suggesting that its staged approach minimizes L1 interference, albeit with less effective integration between L1 and L2. OSMT demonstrated the strongest mitigation of negative transfer, with significant reductions in paradigms like coordinate_structure_constraint and dropped_argument, as well as enhanced stability in paradigms where OST showed sharp declines. This highlights OSMT's ability to balance L1 and L2 inputs, effectively reducing syntactic interference through mixed exposure.

While negative transfer dominated in incompatible paradigms, positive transfer effects emerged in the cases of structural alignment between Korean and English. Paradigms such as binding_principle_a and superlative benefited from structural compatibility, where OSMT outperformed both TST and OST due to its integrated L1-L2 training. These results emphasize the importance of mixed training methods in leveraging structural similarities to enhance L2 acquisition. TST, despite reducing negative transfer, does not capitalize on positive transfer opportunities because of its segmented learning stages. OST, while allowing for continuous training, struggles with mitigating L1 interference in paradigms with structural discrepancies. OSMT, by comparison, strikes a balance, reducing negative transfer while amplifying positive effects, making it the most effective method for managing L1 transfer dynamics. These findings underscore the value of OSMT's mixed training approach in addressing the challenges of negative transfer while maximizing the benefits of positive transfer in bilingual neural models.

Table 2. Performance difference between En(L1) and other models

Phenomena	Paradigms	TST	OST	OSMT
agreement_ subject_verb	across_relative_clause	0.000	-0.015*	-0.008
	across_prepositional_phrase	-0.035	-0.045*	-0.046*
	in_question_with_aux	0.034	-0.016*	0.008
	in_simple_question	0.047*	-0.028*	0.000
agreement_ determiner_ noun	across_1_adjective	-0.036*	-0.072*	-0.008
	between_neighbors	-0.060*	-0.118*	0.012
filler-gap	wh_question_subject	0.178	0.176*	0.184
	wh_question_object	-0.335*	-0.509*	-0.301*
island-effects	adjunct_island	0.004	-0.083*	-0.096*
	coordinate_structure_constraint	0.100	0.087	0.125
quantifiers	existential_there	-0.060	-0.083*	-0.041
	superlative	0.070	-0.006	0.105

Table 2. Continued

Phenomena	Paradigms	TST	OST	OSMT
npi_licensing	only_npi_licensor	-0.061	-0.042	0.024
	matrix_question	-0.077	-0.115	-0.092
argument_ structure	transitive	-0.326*	-0.347*	-0.322*
	dropped_argument	0.136*	0.156*	0.098
	swapped_arguments	0.288*	0.271	0.239
irregular	verb	-0.061*	-0.029	-0.035
anaphor_ agreement	pronoun_gender	0.001	0.001	0.002
ellipsis	n_bar	-0.041	-0.038	0.076*
binding	principle_a	0.086	0.011	0.014
case	subjective_pronoun	0.021	0.022	0.022

$^{*}p$<.05

4. Discussion

This study evaluated the impact of different L2 training methods—Two-Stage Training (TST), One-Stage Training (OST), and One-Stage Mixed Training (OSMT)—on cross-linguistic transfer and catastrophic forgetting in neural language models. These experiments provide insight into how each method influences the preservation of L1 knowledge, the dynamics of L2 acquisition, and the syntactic transfer effects between structurally divergent languages like Korean and English.

4.1. Cross-Linguistic Transfer and Catastrophic Forgetting

Unlike TST, both OST and OSMT demonstrated enhanced resistance to catastrophic forgetting, although they achieved this through distinct methods. OST combines L1 and L2 learning into a single stage, unlike TST's separated approach. This unified learning setup in OST led to a more gradual decline in L1 performance during L2 training. By merging L1 and L2 learning into a single phase, OST appears to reduce the severity of catastrophic forgetting. This continuous dual-language exposure helps the model retain a baseline of L1 knowledge as it acquires L2, underscoring the benefits of one-stage training in preserving earlier syntactic knowledge.

OSMT, which mixed L1 and L2 data at each training step, was even more effective in maintaining L1 knowledge during L2 acquisition. Notably, OSMT even showed slight gains in L1 performance throughout L2 training, suggesting that this balanced, alternating input strategy not only stabilizes but may also reinforce previously learned structures. By continually switching between English and Korean inputs, OSMT better addresses the plasticity-stability dilemma than OST, facilitating new L2 learning while protecting L1 knowledge. This continuous, balanced integration highlights OSMT's strength in fostering bilingual learning with minimal interference.

In Experiment 2, TST demonstrated the quickest increase in L2 English performance, mirroring the rapid progress typical of initial L1 learning. This fast improvement in English accuracy, similar to that of a monolingual English model, suggests that the model's L2 learning was largely unaffected by prior L1 Korean knowledge—a strong indicator of catastrophic forgetting, where L1 knowledge fails to carry over into L2 acquisition. Conversely, OST and OSMT showed more incremental gains in L2 English, suggesting ongoing influence from L1 Korean, which helped guide and stabilize the learning process. These results from Experiment 2 align with those of Experiment 1, where continued L1 exposure in OST and OSMT contributed to mitigating catastrophic forgetting, underscoring the value of integrated training for maintaining bilingual knowledge retention and acquisition.

4.2. Syntactic Transfer Effects between L1 and L2

The experiments also highlighted the syntactic transfer effects, particularly the differences in transfer dynamics across training methods. In Experiment 1, paradigm-specific analyses revealed nuanced interactions between English (L1) and Korean (L2) syntactic knowledge. Transfer effects were observed in specific syntactic paradigms, with both positive and negative impacts depending on structural compatibility between the languages.

Negative transfer was most evident in paradigms like agreement_subject_verb, where Korean lacks direct equivalents to English grammatical structures. TST, by segmenting L1 and L2 training, showed pronounced negative transfer in these areas, as the lack of integrated learning resulted in limited cross-linguistic adaptability. In

OST and OSMT, however, these effects were mitigated, with OSMT showing the least interference. The mixed exposure in OSMT allowed the model to gradually adapt to L2 structures, particularly in incompatible paradigms, achieving a stable balance between L1 and L2. This suggests that OSMT's continuous training method enables the model to reconcile divergent syntactic rules more effectively than separate or sequential learning.

Positive transfer effects, on the other hand, were observed in paradigms with structural alignment, such as binding_principle_a. Here, both Korean and English share compatible anaphoric structures, allowing the model to leverage prior L1 syntactic knowledge to enhance L2 learning. OSMT, with its integrated L1-L2 training, showed the strongest positive transfer effects, indicating that mixed data input allows the model to exploit structural similarities more effectively. This finding underscores the importance of training approaches that accommodate structural alignment in bilingual language models, as they enable models to generalize knowledge across languages with overlapping features.

In Experiment 2, similar patterns were evident. The rapid improvement in English syntactic accuracy in TST was again achieved with minimal influence from L1 Korean, highlighting a reduced impact of cross-linguistic transfer when training stages are separated. OSMT, however, consistently demonstrated balanced transfer effects, achieving a stable learning trajectory even in syntactically incompatible paradigms. These results indicate that a mixed L1-L2 training approach is more conducive to sustainable bilingual learning.

4.3. Broader Implications for Second Language Acquisition Theory

The findings from this study offer an empirical basis for exploring the implications of two key perspectives in second language acquisition, the Representational Deficit Approach (RDA) and the Computational Difficulty Approach (CDA). According to the RDA, adults struggle with L2 because of inherent knowledge gaps, implying a lasting deficiency in L2 acquisition (Hawkins & Chan, 1997). In contrast, the CDA attributes these challenges to performance issues, pointing to the substantial role of learners' first language (L1) as a factor (Haznedar & Schwartz, 1997; Prévost & White, 2000). The fact that OSMT's performance either stabilizes or improves over time, despite the introduction of L2 Korean, indicates the model's ability to integrate syntactic knowledge from multiple languages. This outcome supports the CDA view, suggesting that sufficient exposure and computational resources can help models overcome initial L2 learning challenges. The observed reduction in negative transfer effects with OSMT underscores the idea that computational limitations, rather than intrinsic knowledge deficits, are primarily responsible for L2 difficulties.

4.4. Limitations and Future Research Directions

While this study provides valuable insights into the dynamics of cross-linguistic transfer and catastrophic forgetting in neural language models, several limitations should be noted. The study's focus on two languages with typologically distinct structures (Korean and English) provides a robust test case but may limit the generalizability of findings across other language pairs. Future research should explore a broader range of languages, including typologically similar pairs, to assess whether one stage training methods like OST and OSMT offer comparable benefits across varying degrees of structural similarity.

Another limitation of this study lies in its methodology for analyzing L1 transfer. While the study compares TST, OST, and OSMT with monolingual language model to control for L1 influence on L2 learning, this approach does not fully align with the methodological rigor outlined by Jarvis (2000). According to him, identifying L1 transfer requires satisfying three conditions: (a) intra-L1-group similarities, (b) inter-L1-group differences, and (c) L1-IL performance similarities, which necessitate multiple L1 groups for comparison. Although many studies have discussed L1 effects without strictly adhering to this framework, the absence of multiple L1 groups in this study limits its ability to definitively attribute observed effects to L1 transfer. Future studies need to consider the methodological constraint in interpreting the findings.

Additionally, this study primarily focused on syntactic knowledge as measured by the Zorro test set. Future work could expand this focus to include semantic and pragmatic knowledge, examining whether mixed training approaches similarly benefit other language dimensions. Furthermore, investigating the potential impact of different data sizes and composition ratios within mixed training methods could provide more granular insights into the optimal balance between L1 and L2 input for cross-linguistic transfer.

Finally, while the experiments employed models with relatively limited data exposure compared to real-world neural models, scaling these findings to larger architectures and datasets could reveal further insights. Future studies should examine whether the benefits of OSMT in managing catastrophic forgetting and transfer effects hold under more extensive data conditions, which are typical for state-of-the-art language models.

5. Conclusion

In conclusion, this study demonstrates that one stage training methods like OST are effective in retaining L1 knowledge while facilitating L2 acquisition, particularly in managing catastrophic forgetting and cross-linguistic transfer. Furthermore, we show that mixed training methods like OSMT amplify this effect. By continuously integrating L1 and L2 inputs, OSMT achieves a balance between language retention and adaptation, offering a robust approach for bilingual neural models. These findings have significant implications for the design of multilingual language models with developmentally plausible data volume, supporting the use of mixed training to enhance cross-linguistic transfer in a way that mirrors human-like language learning processes. Through continued exploration of training methods and linguistic structures, future research can further optimize bilingual language models, contributing to a deeper understanding of cross-linguistic effects and more versatile and human-like NLP systems.

References

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., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *arXiv preprint* arXiv:1809.05053.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179-211.

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

Hawkins, R., & Chan, C. Y. (1997). The partial availability of Universal Grammar in second language acquisition: The ‘failed functional features hypothesis’. *Second Language Research*, *13*(3), 187-226.

Haznedar, B., & Schwartz, B. D. (1997). Are there optional infinitives in child L2 acquisition. *Proceedings of the 21st Annual Boston University Conference on Language Development*, *21*, 257-268.

Huebner, P. A., Sulem, E., Fisher, C., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624-646.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language Learning, 50*(2), 245-309.

Jarvis, S., & Pavlenko, A. (2007). *Crosslinguistic influence in language and cognition*. Routledge.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60-99.

Johnson, J. S., & Newport, E. L. (1991). Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition*, *39*(3), 215-258.

Kaushik, P., Gain, A., Kortylewski, A., & Yuille, A. (2021). Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *arXiv preprint arXiv:2102.11343.*

Kemker, R., McClure, M., Abitino, A., Hayes, T. L., & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521-3526.

Koo, K. W., Lee, J. M., & Park, M. K. (2024). Investigating syntactic interference effects in neural language models for second language acquisition. *English Language and Linguistics*, *30*(1), 69-88.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 2935-2947.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? *arXiv preprint arXiv:2005.00955.*

Lopez-Paz, D., & Ranzato, M. A. (2017). Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems, 30.*

McManus, K. (2021). *Crosslinguistic influence and second language learning*. Routledge.

National Institute of Korean Language. (2020). Modu corpus: Open Korean language corpus. *National Institute of Korean Language*. https://corpus.korean.go.kr/

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

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

Prévost, P., & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, *16*(2), 103-133.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007-1028.

Van Schijndel, M. (2019). Quantity doesn't buy quality syntax with neural language models. *arXiv preprint arXiv:1909.00111.*

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. (2022). What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, 17-60, CRC Press.

Warstadt, A., Williams, A., Liu, H., Warstadt, H., Fish, J., & Bowman, S. R. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning.*

Yadavalli, A., Yadavalli, A., & Tobin, V. (2023). SLABERT talk pretty one day: Modeling second language acquisition with BERT. *arXiv preprint arXiv:2305.19589.*

Zhang, Y., Warstadt, A., Li, H. S., & Bowman, S. R. (2020). When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946.*

Jaemin Lee
Graduate Student
Department of English Language and Literature
Dongguk University
30 Pildong-ro 1-gil Jung-gu Seoul, 04620, Korea
E-mail: whd7987@gmail.com

Jeong-Ah Shin
Professor
Department of English Language and Literature
Dongguk University
30 Pildong-ro 1-gil Jung-gu Seoul, 04620, Korea
E-mail: jashin@dongguk.edu

Received: November 14, 2024
Revised version received: December 15, 2024
Accepted: December 16, 2024