Evaluating Syntactic Generalization in Transformer-Based Models Using Korean Honorific Agreement

Nayoung Kwon[♠] **Seongmin Mun**[◊]
[♠]University of Oregon [◊]Kyungpook National University
nkwon@uoregon.edu seongminmun@knu.ac.kr

Abstract

This paper examines whether Transformer-based language models can generalize over abstract syntactic structures in low-resource languages. Focusing on Korean subject–verb honorific agreement, we evaluate KoBERT and KoGPT-2 using classification and attention analyses before and after fine-tuning. Results show that KoGPT-2, especially after fine-tuning, outperforms KoBERT in handling structurally complex constructions. Attention analyses reveal that KoGPT-2 shows more syntactic alignment than KoBERT, although inconsistently. Both models remain susceptible to interference from honorific attractors. These findings highlight key differences between autoregressive and masked LMs in syntactic generalization and show that attention may reflect syntactic structure but not reliably indicate grammatical competence.

1 Introduction

Transformer-based language models have achieved strong performance across a wide range of natural language processing tasks. Although growing evidence suggests that these models implicitly encode aspects of syntactic knowledge (Clark et al., 2019; Hewitt & Manning, 2019; Lin, Tan, & Frank, 2019; Wilcox, Futrell, & Levy, 2024), it remains unclear to what extent they can track syntax-sensitive dependencies and generalize over abstract syntactic structures independently of lexical content. Most work has focused on high-resource languages like English, while low-resource languages such as Korean—making up less than 1% of web content—pose unique challenges due to limited data. This study uses Korean to evaluate the syntactic knowledge encoded in Transformer-based models, focusing specifically on subject–verb honorific agreement.

Although sentences appear linear on the surface, linguistic theory has shown language is fundamentally hierarchical. This raises the question of whether language models can capture such structural information beyond surface patterns. A common strategy for probing syntactic generalization involves agreement phenomena, where two linguistic elements must match in morphosyntactic features. For instance, Goldberg (2019) showed BERT (Devlin et al., 2019) performs well on subject–verb agreement in English, suggesting that Transformer-based models can track syntactic dependencies. Bacon and Regier (2019) replicated these findings with accuracy above 90% across 26 languages. Their results further showed that model performance declines in the presence of agreement attractors and longer dependencies. More recently, Lasri, Lenci, and Poibeau (2022) demonstrated that Transformer-based models' generalization abilities are not fully lexically independent, particularly when processing sentences with attractors. Chaves and Richter (2021) similarly observed that while BERT encodes rich syntactic representations, it often relies on shallow heuristics (see also Wu & Dredze, 2020), in contrast to GPT-2 (Radford, 2019), which exhibited more informed behavior (Chang & Bergen, 2024). Taken together, these findings highlight the need for further investigation into the syntactic generalization capacities of language models, especially in typologically diverse and low-resource languages.

To this end, the present study evaluates the syntactic generalization abilities of KoBERT (Jeon, Lee, & Park, 2019) (92 million parameters) and KoGPT-2 (Jeon, 2021) (125 million parameters), using subject–verb honorific agreement in Korean. By evaluating model behavior before and after fine-tuning, we examine whether syntactic knowledge can emerge in the absence of explicit

syntactic modules, and how such knowledge interacts with attention mechanisms. This study contributes to our understanding of how neural language models engage with syntactic structure in a low-resource language, with implications for multilingual NLP and model interpretability.

2 Subject–Verb Honorific Agreement in Korean

Korean is an agglutinative SOV language in which each morpheme typically encodes a distinct grammatical function. For example, the honorific marker *-si-* attaches to a verb to signal respect for the subject (1). While the use of *-si-* is optional when the subject is honorifiable (2), it becomes ungrammatical with an unhonorifiable subject (3) (Sohn, 2001). Such violations elicit a P600 response (Kwon & Sturt, 2024), similar to effects reported for number and person agreement violations in English (Osterhout & Mobley, 1995) and Spanish (Barber & Carreiras, 2003). This suggests that Korean honorific agreement is processed as syntactic information, akin to other agreement phenomena.

(1) Honorifiable subject with *-si-*
emenim-i wu-si-ess-ta
mother-nom cry-HON-past-decl
'Mother cried.'

(2) Honorifiable subject without *-si-*
emenim-i wul-ess-ta
mother-nom cry-past-decl
'Mother cried.'

(3) Unhonorifiable subject with *-si-*
*kkoma-ka wu-si-ess-ta
kid-nom cry-HON-past-decl
'The kid cried.'

That is, subject–verb honorific agreement in Korean is conditioned by the syntactic accessibility of the subject (Yoon, 2009). While the use of the honorific marker *-si-* reflects social and pragmatic features such as respect, its grammaticality depends on whether the subject structurally licenses agreement. Thus, as an optional but structurally constrained phenomenon, it provides a unique testing ground for determining whether language models can acquire abstract syntactic dependencies without categorical surface cues.

Thus, this study uses subject–verb honorific agreement to test whether Transformer-based language models can generalize over abstract syntactic structures in Korean, a low-resource language. We also evaluate whether fine-tuning on honorific agreement data enhances models' sensitivity to syntactic dependencies. To this end, we constructed sentences with two syntactic configurations illustrated in English in (4) (Chomsky, 1981; Kwon & Polinsky, 2006): in NP1 control, the subject of the embedded verb (underlined) is the main clause subject, NP1; in NP2 control, it is a direct or indirect object, NP2.

(4) NP1-NOM NP2-to $\underline{\text{go}}_{\text{.emb}}$ $\text{told}_{\text{main}}$
NP1 control: 'NP1$_i$ told NP2 that he$_i$ went.'
NP2 control: 'NP1 told NP2$_i$ PRO$_i$ to go.'

To test whether KoBERT and KoGPT-2 have learned syntactic representations, we use subject–verb honorific agreement as a diagnostic via a classification task and self-attention analysis.

3 Experiment

3.1 Datasets

The dataset included the sentence types illustrated in (4), along with matched ungrammatical counterparts in which the honorific verb appears with a structurally licit but non-honorifiable subject. We also varied the honorific features of structurally illicit potential subjects to test for interference. This yielded all four NP1–NP2 feature combinations (H–H, H–NH, NH–H, NH–NH) across both NP1- and NP2-control types. Since the embedded verb is always honorific, NP1-control sentences require NP1 to be honorific, and NP2-control sentences require NP2—regardless of the other noun's features. This design isolates structural understanding from lexical honorific effects. For clarity, sample sentences are shown in English in (5) and (6), although the study was run in Korean. Note that asterisks (*) indicate grammatical violations in Korean (Sohn, 2001), as summarized in Table 1.

(5) NP1 control: The *kid$_i$/teacher$_i$ told the teacher/kid that __$_i$ closed (honorific) the door.

(6) NP2 control: The kid/teacher told the teacher$_i$/*kid$_i$ __$_i$ to close (honorific) the door.

Honorific features		Control type	
NP1	NP2	NP1	NP2
H	H	✓	✓
	NH	✓	✗
NH	H	✗	✓
	NH	✗	✗

Table 1: Grammatical acceptability of dataset

Reflecting naturally occurring distributional patterns in Korean, the dataset contained 1,336 NP1-control sentences and 5,304 NP2-control sentences, which were used for training and evaluation. We split the data into training (90%) and test (10%) sets.

3.2 Classification task analysis

Following previous studies (e.g., McCormick, 2019; Vázquez, 2020; Wu, 2019), we implemented a binary classification task to evaluate whether the models could distinguish grammatically acceptable from unacceptable sentences based on honorific agreement. For KoBERT, we used standard embedding techniques: token, position, and segment embeddings (Devlin et al., 2019). Each sentence was tokenized using the KoBERT tokenizer, truncated or padded to a maximum length of 256 tokens. Tokens were indexed based on the KoBERT vocabulary. Segment embeddings were assigned as binary indicators (0 or 1) to distinguish between sentence segments. Grammaticality labels were stored separately for use in supervised training. Training parameters were set as follows: batch size = 16, number of epochs = 30, random seed = 42, maximum sequence length = 256, epsilon = 8e–8, and learning rate = 0.0001. We fine-tuned KoBERT (Jeon, Lee, and Park 2019) on our dataset using the BertForSequenceClassification class from Huggingface's Transformers library (Wolf, 2019). Following training, we evaluated the model's performance on previously unseen sentences from the test set.

The KoGPT-2 training procedure followed a similar setup, with notable differences in input handling and model architecture. Unlike BERT, GPT-2 uses bytepair encoding (BPE) instead of WordPiece tokenization. In terms of training objectives, BERT is trained using masked language modeling and next-sentence prediction, whereas GPT-2 is trained using a causal (left-to-right) language modeling objective. Additionally, BERT processes input bidirectionally, while GPT-2 processes text unidirectionally, from left to right, without relying on special [CLS] or [SEP] tokens. We used KoGPT-2-base-v2 (Jeon, 2021), implemented using the GPT2ForSequenceClassification and PreTrainedTokenizerFast classes from the Transformers library (Wolf, 2019).

We fine-tuned both models for 30 epochs on the subject–verb honorific agreement dataset and evaluated their classification accuracy and attention alignment before and after training. This setup allowed us to assess their ability to generalize over syntax-sensitive dependencies and examine whether fine-tuning enhances syntactic sensitivity in low-resource settings.

3.3 Attention patterns analysis

Prior work has shown that attention maps in models like BERT can capture meaningful linguistic patterns (see Clark et al., 2019; DeRose, Wang, & Berger 2020; Vig, 2019; Park et al., 2019; for different views, see Jain &Wallace, 2019; Mohankumar et al., 2020; Serrano & Smith, 2019; Thorne et al., 2019). Thus, to investigate how the models represent syntactic dependencies in Korean, we conducted an attention analysis focusing on self-attention weights originating from the critical first verb (VERB1) in each sentence. This verb corresponds to the embedded predicate, where honorific agreement is morphologically marked by *-si-*. The goal of the analysis was to determine whether the model allocates greater attention to the syntactically appropriate subject—either the main clause subject (NP1) in NP1-control sentences or a structurally lower NP (NP2) in NP2-control sentences—when computing the contextual representation of VERB1. We extracted attention weights from VERB1 to the two potential subjects, NP1 and NP2. These values reflect the extent to which VERB1 attends to information provided by each NP, serving as an indirect measure of which constituents the model treats as syntactically or semantically relevant.

Because KoBERT and KoGPT-2 differ in architecture and implementation, raw attention weights are not directly comparable across models. To allow for meaningful comparison, we analyzed normalized attention ratios, which capture the relative distribution of attention across potential antecedents (NP1 vs. NP2). Attention weights were extracted during the models' evaluation of sentence grammaticality at both Epoch 1 (pre-trained) and

Epoch 30 (fine-tuned). For each trial, we computed the attention ratio for NP1 as the proportion of attention allocated to NP1 relative to the total attention directed to NP1 and NP2 (i.e., NP1 Ratio = NP1 Attention / [NP1 Attention + NP2 Attention] × 100), and likewise for NP2. This normalized, directional metric allows us to assess attention patterns independently of differences in absolute attention scale. To quantify relative attentional preference, we then calculated an attention difference score for each trial by subtracting the NP2 ratio from the NP1 ratio (i.e., NP1 Ratio – NP2 Ratio). Accordingly, positive values indicate a preference for NP1, while negative values reflect greater attention to NP2.

To explore how attention patterns relate to classification outcomes, we selected 400 sentences that were correctly classified by both models and 56 that were misclassified by both. These subsets formed the basis of our attention analysis. Transformer models compute attention matrices over 12 layers and 12 heads based on tokenized inputs. Because tokenization significantly affects the model's interpretation of sentence structure, it plays a critical role in attention analysis. In this study, we used a syllable-based tokenizer, which occasionally led to token splitting inconsistencies due to whitespace handling. To enhance interpretability while preserving the basic clause structure [NP1 NP2 VERB1 VERB2], we adopted a modified version of the method proposed by Mun and Shin (2025), as illustrated in Appendix A.

3.4 Statistical analyses[1]

Classification accuracy was analyzed using generalized linear mixed-effects models (GLMER) with a binomial link, implemented in R 4.2.1 (R Core Team, 2024) via the lme4 package (Bates et al., 2015, v1.1-31). P-values were computed using *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017, version 3.1-3). Attention data were analyzed using linear mixed-effects regression (LMER). Classification accuracy models included four fixed effects: Model (KoBERT vs. KoGPT-2), Control Type (NP1 vs. NP2), and the Honorific features of NP1 and NP2 (H vs. NH), along with all interactions. Attention models included Model, Control Type, and Classification Accuracy (correct vs. incorrect) as fixed effects. To evaluate the impact of fine-tuning on syntactic sensitivity, we compared Epoch 1 (pre-trained) and Epoch 30 (fine-tuned) performance using models with Model, Epoch, and Control Type as predictors. All predictors were sum-coded. Random intercepts for sentence were included to account for trial-level variability. Random intercepts for sentence were included, and maximal random-effects structures (Barr et al., 2013) were simplified for convergence. Holm-adjusted pairwise comparisons were conducted using the *emmeans* package (Lenth, 2024, v1.8.4-1).

3.5 Results and Discussion

Classification Accuracy Before and After Fine-tuning (Epoch 1 vs. Epoch 30). Appendix B presents the grammatical acceptability classification accuracies for KoBERT and KoGPT-2. Corresponding statistical results are in Appendix C. The analysis showed main effects of Model and Control Type: KoGPT-2 (M = 89.7%) significantly outperformed KoBERT (M = 81.1%), and NP2-control sentences (M = 87.6%) were classified more accurately than NP1-control sentences (M = 76.8%). Honorific features of NP1 and NP2 also had main effects, indicating sensitivity to subject–verb honorific agreement. A significant Control Type × NP1 × NP2 interaction revealed that classification accuracy varied by agreement configuration. When agreement was disrupted by a feature-matching attractor—e.g., NP1-control with non-honorific NP1 and honorific NP2—accuracy dropped sharply (KoBERT: 45.3%; KoGPT-2: 52.3%). In contrast, when the attractor did not linearly intervene—as in NP2-control with honorific NP1 and non-honorific NP2—KoGPT-2 maintained high accuracy (92.4%), correctly identifying these sentences as ungrammatical, whereas KoBERT's accuracy was substantially lower (76.7%). KoGPT-2 significantly outperformed KoBERT across conditions (ps $< .03$), except in NP1-control sentences with honorifiable NP1 and non-honorifiable NP2, where KoBERT (81.3%) outperformed KoGPT-2 (69.5%) ($p < .0001$).

These findings suggest that language models' performance declines with increasing agreement distance and feature-matching attractors that intervene in subject–verb dependencies (Bacon &

[1] All data and analysis code are available at the following link: https://osf.io/fw82j/?view_only=56b9d1dc865e453086dfbc8957fee340

Regier, 2019; Ryu & Lewis, 2021; Lakretz et al., 2022), paralleling patterns in human sentence processing (e.g., Kwon & Sturt 2016, 2019). The results were also consistent with previous studies showing that autoregressive models like GPT-2 are more robust to such interference and better at maintaining long-distance syntactic dependencies, whereas masked language models like BERT are more susceptible to feature-based interference from structurally irrelevant attractors (Chaves & Richter, 2021; Lasri, Lenci, & Poibeau, 2022).

We next examined whether fine-tuning improves the models' sensitivity to syntactic structure. Appendix D presents the classification accuracy of fine-tuned KoBERT and KoGPT-2, by control type and honorific features of NP1 and NP2. Appendix E presents the corresponding statistics, and Figure 1 visualizes aggregated results by model, epoch, and control type.

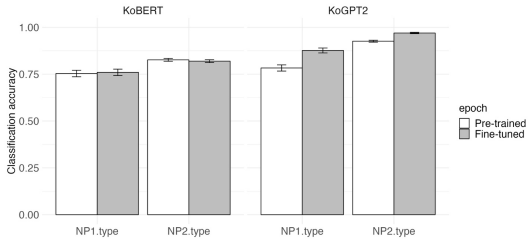


Figure 1. Classification accuracy aggregated across honorific conditions, grouped by model, control type, and training state (pre-trained vs. fine-tuned: Epoch 1 vs. Epoch 30)

Analysis revealed significant main effects of Model, Control type, and Epoch, indicating overall higher accuracy for KoGPT-2 (92.4%) than KoBERT (80.9%), and for NP2-control (88.5%) over NP1-control (79.3%) sentences. Accuracy also improved after fine-tuning (88.0%) compared to the pre-trained state (85.4%). These main effects were qualified by significant interactions, including Model × Control type and Epoch × Model and a three-way Model × Epoch × Control type interaction. Holm-adjusted pairwise comparisons showed that KoGPT-2 exhibited substantial gains after fine-tuning for both NP1-control ($z = -8.21$, $p < .001$) and NP2-control sentences ($z = -9.95$, $p < .001$). In contrast, KoBERT showed no significant improvement for NP1-control ($z = 1.20$, $p = .23$) and even declined in NP2-control accuracy ($z = 4.01$, $p < .001$).

During pre-training, both KoBERT and KoGPT-2 exhibited high error rates when a feature-matching attractor linearly intervened in subject–verb honorific agreement (i.e., NP1-control sentences with a non-honorifiable NP1 and an honorifiable NP2). After fine-tuning, both models showed improved classification accuracy for this construction (KoBERT: 45.3% → 61.8%; KoGPT-2: 52.4% → 72.8%). To evaluate whether fine-tuning reduced attractor interference, we conducted a follow-up analysis. Focusing on NP1-control sentences, we compared conditions in which the honorific features of NP1 and NP2 differed (H–NH and NH–H) to those in which they matched (H–H and NH–NH). If models adhered strictly to syntactic structure, performance should be unaffected by the features of an illicit subject. However, Welch's two-sample *t*-tests revealed that mismatched conditions significantly reduced classification accuracy for both models: H–NH vs. H–H (KoBERT: $t(1089.3) = 5.60$, $p < .001$; KoGPT-2: $t(679.16) = 10.87$, $p < .001$), and NH–H vs. NH–NH (KoBERT: $t(1075.9) = -6.57$, $p < .001$; KoGPT-2: $t(635.37) = -14.25$, $p < .001$). These findings suggest that although fine-tuning substantially improved KoGPT-2's overall accuracy and structural sensitivity, both models remain vulnerable to honorific feature interference in structurally complex NP1-control configurations.

Overall, the results suggest that KoGPT-2 benefits substantially from fine-tuning, showing improved classification accuracy across both control types. In contrast, KoBERT appears less responsive to fine-tuning and even declined in NP2-control performance. This divergence may reflect architectural differences in how the two models encode syntactic dependencies, with KoGPT-2's autoregressive design better supporting structural sensitivity. Nonetheless, both models remain vulnerable to honorific feature interference, especially in NP1-control sentences where structurally irrelevant NPs disrupt subject–verb agreement.

To further examine how syntactic information is internally represented, we next analyze the models' attention patterns during the classification task.

Attention Patterns Before and After Fine-Tuning (Epoch 1 vs. Epoch 30). The attention analysis included 456 trials from the classification task that were either correctly (n = 400) or incorrectly (n = 56) classified by both models. This set comprised 104 NP1-control and 352 NP2-control sentences. By comparing correct and incorrect classifications, we aimed to explore how attention patterns relate to model performance and to identify structurally

relevant attention cues that may support accurate classification.

Because absolute attention weights vary between GPT-2 and BERT due to differences in architecture and implementation, we focused on relative attention distributions. Specifically, we analyzed preference attention ratios, which quantify how attention is distributed between NP1 and NP2, independent of model scale (see Section 3.3). In our analysis, positive values indicate a preference for NP1, while negative values reflect greater attention to NP2. Accordingly, under this metric, a structurally aligned model should yield more positive scores for NP1-control sentences and more negative scores for NP2-control sentences.

Figure 2 presents the attention difference scores by model and classification accuracy for pre-trained KoBERT and KoGPT-2. The statistical analysis results are presented in Appendix F.

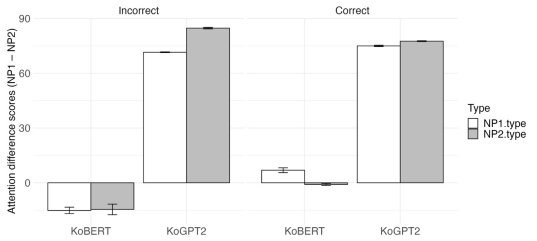


Figure 2. Attention difference scores by model and classification accuracy for KoBERT and KoGPT-2 in their pre-trained states.

The analysis revealed significant main effects of Model and Classification Accuracy: KoGPT-2 had higher attention difference scores (M = 76.8, SE = 0.55) than KoBERT (M = −1.6, SE = 1.48), suggesting that KoGPT-2 allocated more attention to NP1 than KoBERT. Correct trials also yielded higher scores (M = 38.7, SE = 2.84) than incorrect ones (M = 0.29, SE = 8.89). These main effects were moderated by significant interactions: Model × Classification accuracy and Model × Control type. Holm-corrected pairwise comparisons indicated that KoBERT allocated more attention to NP2 in incorrect trials than in correct ones ($t(192) = -4.45$, $p < .0001$), whereas KoGPT-2 did not show a comparable difference ($t(192) = 0.45$, n.s.). In addition, KoGPT-2 allocated more attention to NP1 in NP2-control than NP1-control sentences ($t(192) = -1.97$, $p = .05$). In contrast, KoBERT did not exhibit such an asymmetry ($t(192) = 0.90$, n.s.).

These findings suggest that, in the pre-trained state, attention allocation patterns do not consistently reflect syntactic roles, despite decent classification accuracy. KoGPT-2 directed significantly more attention to NP1 in NP2-control sentences, where NP2 is the syntactically appropriate subject. This indicates a misalignment between attention and the underlying grammatical dependencies. KoBERT's attention favored NP2 in incorrect trials, likely reflecting interference rather than accurate subject identification. Overall, these results highlight that attention distributions in pre-trained models may not consistently indicate syntactic understanding. This aligns with prior work demonstrating weak links between attention weights and model performance or reasoning processes (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019).

Having established baseline attention patterns in the pre-trained models, we next examined how fine-tuning affected attention allocation in KoBERT and KoGPT-2. Using the same statistical models, we evaluated whether fine-tuning improved the alignment between attention and syntactic roles. Figure 3 presents the attention difference scores (NP1 – NP2) by model, classification accuracy, and control type for the fine-tuned models. The corresponding statistical analysis results are shown in Appendix G.

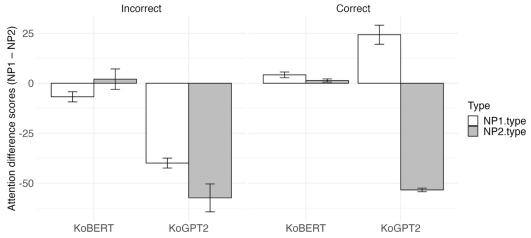


Figure 3. Attention difference scores by model and classification accuracy for fine-tuned KoBERT and KoGPT-2

The analysis revealed significant main effects of Model and Classification accuracy. KoGPT-2 showed lower attention difference scores (M = −41.9, SE = 3.65) than KoBERT (M = 0.95, SE = 1.91), and correctly classified trials yielded higher scores (M = −19.9, SE = 2.72) than incorrectly classified ones (M = −24.3, SE = 6.11). A significant main effect of Control type also emerged, with NP1-control sentences eliciting higher scores (M = −1.69, SE = 5.56) than NP2-control sentences (M = −26.1, SE = 2.66). These main effects were qualified by significant interactions: Model × Control type, Model × Classification accuracy, Control type ×

Classification accuracy, and a three-way interaction of Model × Control type × Classification accuracy. Holm-corrected comparisons showed that KoGPT-2 allocated significantly more attention to NP1—the correct subject—in correctly classified NP1-control sentences than in incorrect ones (t = −6.40, p < .0001), indicating greater syntactic alignment. No such effect was observed for KoGPT-2 in NP2-control sentences (t = −0.27, p = .789) or for KoBERT in either condition (ps > .78).

These results after fine-tuning shed light on the interpretability of attention in Transformer models, particularly regarding their sensitivity to syntactic roles. The significant three-way interaction among model type, control type, and classification accuracy suggests that attention allocation is not uniformly predictive of classification performance but is modulated by both structural configuration and model architecture. Notably, KoGPT-2 showed a strong link between attention and classification accuracy in NP1-control sentences after fine-tuning, suggesting improved syntactic alignment. This effect was absent in NP2-control sentences, likely due to the small number of classification errors (n = 6), which limited statistical power. In contrast, KoBERT's attention was unaffected by classification accuracy following fine-tuning, aligning with its limited performance gains during training. This suggests that KoBERT's attention patterns may be less sensitive to syntactic structure or more weakly coupled with task success, echoing previous findings that attention weights do not always reflect meaningful model behavior (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019).

Taken together, these findings highlight the importance of evaluating attention behavior in relation to both model architecture and syntactic structure. They also caution against interpreting attention weights as direct indicators of linguistic competence: while attention can reflect syntactic alignment in some cases, this is not a reliable property across models or sentence types. KoBERT's lack of attention modulation despite high classification accuracy raises questions about the interpretability of attention in masked language models. We return to this issue in the general discussion.

4 General Discussion and Conclusion

This study aimed to investigate whether Transformer-based language models can encode abstract syntactic structures, even in the absence of an explicit layer dedicated to syntactic representation. We addressed this question using Korean, a low-resource language that differs typologically from widely studied English, by examining model behavior through the lens of subject–verb honorific agreement. Although honorifics may appear socio-pragmatic in nature, the use of the honorific suffix *-si-* requires a licensing subject with matching features in a structurally appropriate position. Accordingly, -si- honorific agreement provides a unique testing ground for evaluating whether language models can acquire abstract syntactic dependencies in the absence of categorical surface cues. To this end, we employed both a grammaticality classification task and a self-attention analysis to evaluate the performance of two Korean-language models, KoBERT and KoGPT-2, in both their pre-trained states and after fine-tuning (i.e., before and after exposure to honorific agreement patterns in Korean). The training and classification datasets included sentences featuring subject–verb honorific agreement that varied in control type and in the honorific features of the potential subject NPs.

Our results offer three key contributions to the study of syntactic generalization in neural language models. First, although both models achieved relatively strong classification performance—suggesting some degree of syntactic understanding consistent with prior work (Clark et al., 2019; Hewitt & Manning, 2019; Lin, Tan, & Frank, 2019)—KoGPT-2 consistently outperformed KoBERT and responded more robustly to fine-tuning (Chaves & Richter, 2021). Even in their pre-trained states, KoGPT-2 generally outperformed KoBERT. While KoBERT showed only modest gains—or even declines—in performance after fine-tuning, KoGPT-2 improved significantly across both NP1- and NP2-control conditions. Both models struggled with attractor interference, but KoBERT appeared especially susceptible. In the syntactically complex NP1-control condition—where the attractor NP2 linearly intervenes between the subject (NP1) and the verb—KoBERT's pre-trained accuracy was only 45.3%, compared to 52.4% for KoGPT-2. After fine-tuning, KoGPT-2's performance rose to 72.8%, while

KoBERT reached only 61.8%. These findings support prior claims that autoregressive models like GPT-2 are better equipped to track hierarchical dependencies than masked language models like BERT (Chaves & Richter, 2021; Lasri, Lenci, & Poibeau, 2022). In contrast, BERT may rely more heavily on shallow heuristics, rendering it more susceptible to lexical interference (McCoy, Pavlick, & Linzen, 2019). Importantly, our use of a low-resource language extends these insights beyond high-resource settings yielding critical evidence about models' capacity to generalize over abstract syntactic structure.

Second, our attention analyses contribute empirical evidence to the ongoing debate about the interpretability of attention mechanisms in Transformer-based models. Specifically, our results support previous claims that attention weights do not reliably correspond to linguistic explanations (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019; Zhao et al., 2024) even though alignment does emerge in some cases. For KoGPT-2, fine-tuning led to greater alignment between attention allocation and syntactic roles, but only under specific conditions. In correctly classified NP1-control sentences, KoGPT-2 consistently directed greater attention to NP1, the structurally appropriate subject. In contrast, in misclassified NP1-control trials, attention shifted toward NP2, suggesting that syntactically guided attention supports successful classification. However, this relationship is not guaranteed. For instance, KoBERT showed no such modulation: its attention patterns remained largely insensitive to syntactic structure or classification outcome. Notably, KoBERT's classification accuracy plateaued across training (~81%) and even declined in certain conditions after finetuning—mirroring its lack of syntactic alignment in attention, despite relatively strong overall performance. These results reinforce growing concerns that attention weights, while useful in some contexts, may not consistently reflect underlying grammatical knowledge or task-relevant reasoning. They also underscore the importance of jointly evaluating both model performance and internal interpretability when assessing syntactic generalization in neural language models (Jain and Wallace 2019).

Third, our results underscore the utility of Korean honorific agreement as a rigorous diagnostic for evaluating syntactic sensitivity in language models. Unlike number agreement in English, Korean subject–verb honorific agreement is governed by structural licensing conditions that are not always transparent at the surface level. The optionality and morphosyntactic specificity of -si-honorific agreement allow researchers to probe whether models can move beyond surface-level lexical heuristics and encode deeper grammatical generalizations. On the other hand, the fact that both KoBERT and KoGPT-2 remained susceptible to interference from honorific attractors even after finetuning suggests that fully abstracting over syntactic structure—particularly in constructions involving long-distance dependencies and feature checking—remains a significant challenge for current Transformer architectures.

Taken together, our findings demonstrate that Transformer-based language models can acquire sensitivity to morphosyntactic dependencies, even when these dependencies are tied to socio-pragmatic cues such as honorifics. While both models performed reasonably well in their pre-trained states, fine-tuning—especially for KoGPT-2—led to notable improvements in classification accuracy and more syntactically aligned attention patterns in structurally complex sentences. However, these gains were not uniform. KoGPT-2's attention aligned with syntactic roles only under certain conditions, and KoBERT showed little evidence of syntactically guided attention despite moderate classification accuracy.

These results underscore the limits of using attention as indicators of grammatical knowledge. Attention may sometimes reflect syntactic reasoning, but it does not reliably track structural representations across models or constructions (Jain & Wallace, 2019; Serrano & Smith, 2019; Mohankumar, 2020; Thorne et al., 2019). This study underscores the value of cross-linguistic research, especially with low-resource languages like Korean, whose rich morphology and flexible word order can reveal model limitations obscured in English-centric evaluations (cf. Chang & Bergen, 2024; Wu & Dredze, 2020).

At the same time, it is important to acknowledge certain limitations of our study. In particular, our analysis did not fully address how implementation details might have influenced the results. Factors such as tokenization (e.g., the proportion of [UNK] tokens), differences in model size (KoBERT and KoGPT-2 differ not only in inference type but also in parameter scale), and the characteristics of the

pretraining corpora (e.g., written vs. spoken language styles) could all have contributed to the observed outcomes. To minimize potential confounds, all test sentences were lexically matched across the experimental conditions, with the only differences arising from the honorific features that served as the manipulation. This design should therefore reduce the likelihood that other sentence components drove the observed effects. Nevertheless, we cannot fully exclude this possibility, particularly given differences in the models' pretraining corpora.

Future research should explore whether incorporating explicit syntactic supervision or inductive biases—such as training on treebank-annotated corpora or employing structure-aware architectures—enhances models' ability to generalize robustly and yields more interpretable internal representations, especially across typologically diverse languages.

References

Bacon, G., & Regier, T. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3861–3871).

Barber, H., & Carreiras, M. (2003). Integrating gender and number information in Spanish word pairs: An ERP study. Cortex, 39(3), 465–482.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255–278.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software. https://arxiv.org/abs/1406.5823

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. Computational Linguistics, 50(1), 293–350.

Chaves, R. P., & Richter, S. N. (2021). Look at that! BERT can be easily distracted from paying attention to morphosyntax. In Proceedings of the Society for Computation in Linguistics 2021 (pp. 28–38). Association for Computational Linguistics.

Chomsky, N. (1981). Lectures on government and binding. Foris.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In Proceedings of the 2019 ACL Workshop on Blackbox NLP (pp. 276–286).

DeRose, J. F., Wang, J., & Berger, M. (2020). Attention flows: Analyzing and comparing attention mechanisms in language models. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1160–1170.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 NAACL-HLT (pp. 4171–4186).

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. CoRR, arXiv:1901.05287.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the NAACL-HLT (pp. 4129–4138).

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434–446.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 NAACL-HLT (pp. 3543–3556).

Jeon, H., Lee, D., & Park, J. (2019). Korean BERT pre-trained cased (KoBERT). SK Telecom AI Center. https://github.com/SKTBrain/KoBERT

Jeon, H. (2021). KoGPT2 ver 2.0. Hugging Face. https://huggingface.co/skt/kogpt2

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13), 1–26.

Kwon, N., & Sturt, P. (2016). Attraction effects in honorific agreement in Korean. Frontiers in Psychology, 7, 1302.

Kwon, N., & Sturt, P. (2019). Proximity and same case marking do not increase attraction effect in comprehension: Evidence from eye-tracking experiments in Korean. Frontiers in Psychology, 10, 1320.

Kwon, N., & Sturt, P. (2024). When social hierarchy matters grammatically: Investigation of the processing of honorifics in Korean. Cognition, 251, 105912.

Kwon, N., & Polinsky, M. (2006). Object control in Korean: Structure and processing. Japanese/Korean Linguistics, 15, 249–262.

Lakretz, Y., Desbordes, T., Hupkes, D., & Dehaene, S. (2022). Can transformers process recursive nested constructions, like humans? In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3226–3232).

Lasri, C., Lenci, A., & Poibeau, T. (2022). Syntactic generalization and lexical heuristics in transformer-based language models. In Proceedings of the 60th Annual Meeting of the ACL (pp. 3703–3717).

Lenth, R. V. (2024). emmeans: Estimated marginal means, aka least-squares means (Version 1.10.0) [R package]. https://CRAN.R-project.org/package=emmeans

Lin, Y. C., Tan, Y. C., & Frank, R. (2019). Open sesame: Getting inside BERT's linguistic knowledge. In Proceedings of the 2019 ACL Workshop on Blackbox NLP (pp. 241–253).

McCormick, C. (2019). BERT fine-tuning tutorial with PyTorch. https://mccormickml.com/2019/07/22/BERT-fine-tuning/

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th ACL (pp. 3428–3448).

Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., & Ravindran, B. (2020). Towards transparent and explainable attention models. In Proceedings of the 58th ACL (pp. 4206–4216).

Mun, S., & Shin, G.-H. (2025). Polysemy interpretation and transformer language models: A case of Korean adverbial postposition -(u)lo. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 1555–1561).

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. Journal of Memory and Language, 34(6), 739–773.

Park, C., Na, I., Jo, Y., Shin, S., Yoo, J., Kwon, B. C., Zhao, J., Noh, H., Lee, Y., & Choo, J. (2019). SanVis: Visual analytics for understanding self-attention networks. In IEEE VIS (pp. 146–150).

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (pp. 61–71).

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? In Proceedings of the 57th ACL (pp. 2931–2951).

Sohn, H.-M. (2001). The Korean language. Cambridge University Press.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2019). Generating token-level explanations for natural language inference. In Proceedings of the NAACL-HLT (pp. 963–969).

Vig, J. (2019). Visualizing attention in transformer-based language representation models. arXiv preprint, arXiv:1904.02679.

Vázquez, R., et al. (2020). A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. Computational Linguistics, 46(2), 387–424.

Wilcox, E. G., Futrell, R., & Levy, R. (2024). Using computational models to test syntactic learnability. Linguistic Inquiry, 55(4), 805–848.

Wolf, T., et al. (2019). Transformers: State-of-the-art natural language processing. arXiv preprint, arXiv:1910.03771.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? In Proceedings of the 5th Workshop on Representation Learning for NLP (pp. 120–130).

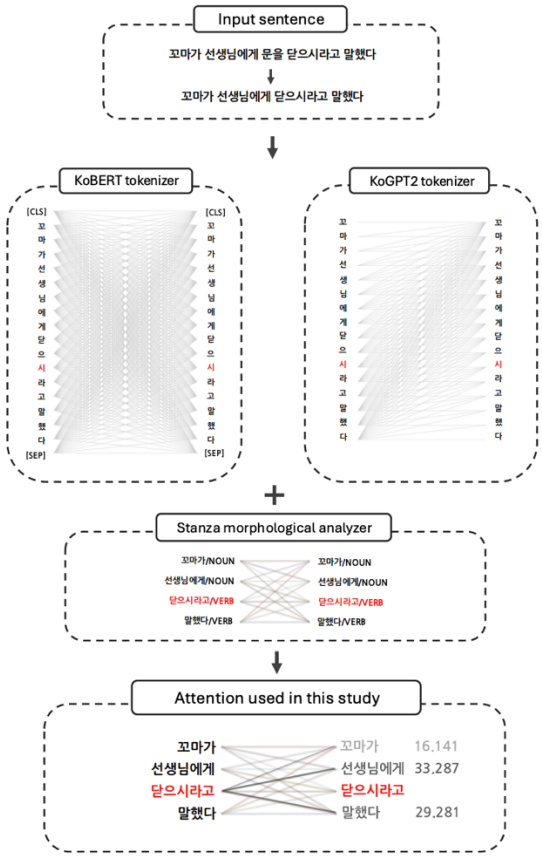
Wu, Y., et al. (2019). A sequential matching framework for multi-turn response selection in retrieval-based chatbots. Computational Linguistics, 45(1), 163–197.

Yoon, J. (2009). The distribution of subject properties in multiple subject constructions. In Japanese/Korean Linguistics (Vol. 19, pp. 64–83).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), Article 20.

Appendices

Appendix A. N/V unit treatment and attention transformation (Example (5))



Appendix B. Classification accuracy results for pre-trained KoBERT and KoGPT-2 on target sentences involving subject–verb honorific agreement. Results are grouped by the honorific features of the two potential subject NPs (NP1 and NP2) and by sentence control type (NP1-control vs. NP2-control).

KoBERT				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	89.1 (0.01)	86.5 (0.01)	81.2 (0.01)
	NH	81.3 (0.02)	76.8 (0.01)	
NH	H	45.3 (0.02)	82.8 (0.01)	
	NH	84.2 (0.02)	84.6 (0.01)	
KoGPT2				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	93 (0.01)	94.3 (0.01)	89.8 (0.01)
	NH	69.5 (0.02)	92.5 (0.01)	
NH	H	52.4 (0.02)	86.6 (0.01)	
	NH	98.4 (0.01)	97 (0.01)	

Appendix C. Linear Mixed Effects model results for classification accuracy by pre-trained KoBERT and KoGPT-2. Coefficients, standard errors, z-values, and p-values are reported for main effects and their interactions. The model included random intercepts for sentence. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***z***	***p***
(Intercept)	2.281	0.05	45.43	.001
NP1	0.153	0.04	3.95	.001
NP2	-0.176	0.05	-3.51	.001
Model	-0.481	0.04	-12.86	.001
Control type	-0.377	0.04	-9.86	.001
NP1:NP2	0.821	0.04	21.16	.001
NP1:Model	0.176	0.04	4.75	.001
NP2:Model	0.159	0.04	4.27	.001
NP1:Type	0.173	0.04	4.53	.001
NP2:Type	-0.249	0.04	-6.53	.001
Model:Type	0.131	0.04	3.52	.001
NP1:NP2: Model	-0.285	0.04	-7.67	.001
NP1:NP2: Type	0.446	0.04	11.71	.001
NP1:Model: Type	0.223	0.04	5.98	.001
NP2:Model: Type	-0.067	0.04	-1.8	0.08
NP1:NP2: Model:Type	-0.145	0.04	-3.91	.001

Appendix D. Classification accuracy results for fine-tuned KoBERT and KoGPT-2 on target sentences involving subject–verb honorific agreement. Results are grouped by the honorific features of the two potential subject NPs (NP1 and NP2) and by sentence control type

KoBERT				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	86.5 (0.01)	86.1 (0.01)	80.8 (0.01)
	NH	74.1 (0.02)	79.2 (0.01)	
NH	H	61.8 (0.02)	84.4 (0.01)	
	NH	79.6 (0.02)	78.5 (0.01)	
KoGPT2				
NP1	NP2	NP1 ctrl	NP2 ctrl	Mean
H	H	98.4 (0.01)	98.8 (0.01)	95.1 (0.01)
	NH	79.6 (0.02)	97.1 (0.01)	
NH	H	72.8 (0.02)	94.4 (0.01)	
	NH	99.4 (0.01)	97.8 (0.01)	

Appendix E. Linear mixed-effects model results for classification accuracy of KoBERT and KoGPT-2 before and after fine-tuning (Epoch 1 vs. Epoch 30). Coefficients, standard errors, *z* and *p*-values are reported for the main effects, as well as for their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***z***	***p*** <
(Intercept)	3.807	0.1	37.56	.001
Epoch	0.725	0.08	8.76	.001
Model	-2.182	0.1	-21.8	.001
Control type	-0.582	0.05	-11.46	.001
Epoch:Model	-0.905	0.08	-10.9	.001
Epoch:Type	-0.056	0.05	-1.08	0.29
Model:Type	0.321	0.05	6.33	.001
Epoch:Type	0.127	0.05	2.46	0.02

Appendix F. Linear mixed-effects model results for attention difference scores in the pre-trained state. Coefficients, standard errors, t-values, and p-values are reported for the main effects and their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***t***	***p*** <
(Intercept)	35.625	1.66	21.41	.001
Model	-41.555	1.12	-37.2	.001
Control type	-1.07	1.66	-0.64	0.52
Classification accuracy	-4.005	1.66	-2.41	0.02
Model:Type	2.881	1.12	2.58	0.02
Model: Accuracy	-4.922	1.12	-4.41	.001
Type: Accuracy	-2.368	1.66	-1.42	0.16
Model:Type: Accuracy	0.285	1.12	0.26	0.8

Appendix G. Linear mixed-effects model results for attention difference scores in the fine-tuned state. Coefficients, standard errors, t-values, and p-values are reported for the main effects and their interactions. Whether a random slope was included for each effect is not shown here but was considered in model fitting.

	Estimate	**SE**	***t***	***p*** <
(Intercept)	-15.72	3.54	-4.45	.001
Model	15.91	2.75	5.79	.001
Control type	11.13	3.54	3.15	0.01
Classification accuracy	-9.81	3.54	-2.78	0.01
Model: Control type	-12.64	2.75	-4.6	.001
Model: Accuracy	7.24	2.75	2.63	0.01
Type: Accuracy	-8.99	3.54	-2.54	0.02
Model:Type: Accuracy	6.08	2.75	2.21	0.03