Can language models process linguistic deference?

Youngdong Cho
Cornell University
yc2544@cornell.edu

Chloe Dokyung Kwon
Cornell Univeristy
dk837@cornell.edu

Abstract

Honorifics are linguistic forms that encode respect toward a socially valued individual or entity. This paper investigates how language models process Korean subject honorifics, which signal the social status of the subject through specific morphological markers. We evaluate a set of language models to determine whether they process honorifics in a human-like way by capturing the socio-pragmatic constraints governing their use, rather than merely relying on surface co-occurrence patterns. Our results indicate a systematic dissociation: models generally succeeded in detecting surface morphosyntactic mismatches, successfully treating unacceptable honorific constructions as less expected. However, models consistently favored overt honorific marking regardless of the subject's social status, suggesting reliance on surface heuristics over genuine pragmatic knowledge. These findings suggest that language models have not fully acquired the socio-pragmatic constraints underlying honorific use, even when extensively trained on Korean text.

1 Introduction

Honorifics are linguistic forms that convey deference towards a socially respected referent, typically a person of superior social standing. This paper aims to see how language models process honorifics, focusing on *subject honorifics*, which reflect the relationship between the speaker and the subject of the sentence. Subject honorifics are realized through special morphological inflections attached to the subject and the verb when the subject is an honorable entity, as illustrated in (1b).[1]

(1) a. Ai-ka pyenci-lul ssu-ess-ta.
child-NOM letter-ACC write-PST-DECL
'The child wrote a letter.'

b. Kyoswunim-**kkeyse** pyenci-lul
professor-**NOM.HON** letter-ACC
ssu-**si**-ess-ta.
write-**HON**-PST-DECL
'The professor wrote a letter.'

Korean is an agglutinative language in which grammatical roles, such as subject and object, are indicated by case markers attached to nouns. The subject is marked with a nominative marker *-i/ka*, and the object is marked with an accusative marker *-ul/lul* as shown in (1a). However, when the subject is an honorable entity such as *kyoswunim* 'professor' in (1b), the plain nominative marker *-i/ka* can be replaced by an honorific nominative marker, and the verb may carry the honorific suffix *-si* after the verbal stem.

At first glance, Korean subject honorifics resemble English subject-verb agreement in that the subject triggers specific morphological inflection on the verb (Linzen et al., 2016). However, the Korean system exhibits a more complex pattern; unlike the obligatory nature of English agreement, honorific markings in Korean remain optional, even though the subject is an honorable entity (Noh et al., 2024). For example, the verbal honorific marker *–si* can be omitted, and the nominative marker can alternatively be realized as the plain marker *–i/ka*. This optionality tends to arise when the speaker perceives themselves as familiar with the referent, or in registers characterized by neutral description, such as news articles.

While language model performance on syntactic dependencies, including filler-gap dependencies (Wilcox et al., 2018), subject-verb agreement (Linzen et al., 2016; Goldberg, 2019; Guarasci et al., 2021), anaphor binding (Hu et al., 2020), and control phenomena (Lee and Schuster, 2022), has been extensively explored in recent years, comparatively little attention has been paid to non-syntactic dependencies that encode pragmatic constraints such as honorifics. Because both morphosyntax

[1] Abbreviations used in this paper are provided in the appendix.

and socio-pragmatic factors condition honorific use, it provides a uniquely revealing testing ground for evaluating whether language models track not only structural dependencies but also the social knowledge that governs their use. The present study addresses this gap by extending Noh et al. (2024). Specifically, whereas their study focused on the presence of the verbal honorific marker *-si*, we employ a full paradigm that also includes the nominative honorific marker *-kkeyse*. Second, while their analyses were limited to Korean-focused encoder models, we also examine multilingual and decoder-based models, enabling a broader assessment of how model architecture and training regime affect the processing of Korean honorifics.

2 Background

2.1 Korean honorific system

There are two main types of honorifics in Korean. First, addressee (or hearer) honorifics indicate the relationship between the speaker and the addressee. When the addressee is socially higher than the speaker, the speaker employs certain sentential endings (e.g., *-yo* or *-supnita*) to express the speaker's deference toward the addressee. On the other hand, subject honorifics reflect the relationship between the speaker and the subject of the sentence. When the subject is an entity considered to be honorable by the speaker, special morphological inflections are attached to the subject (i.e., *-kkeyse*) and the verb (i.e., *-si*), as shown above in (1b).

2.2 Subject honorifics: Human judgment

Unlike honorable subjects, plain subjects do not trigger special morphological inflections. Examples in (2) show that the plain subject *ai* 'child' is not compatible with any honorific markings, where * indicates unacceptability.

(2) a. Ai-ka pyenci-lul ssu-ess-ta.
child-NOM letter-ACC write-PST-DECL

b. *Ai-ka pyenci-lul
child-NOM letter-ACC
ssu-si-ess-ta.
write-HON-PST-DECL

c. *Ai-kkeyse pyenci-lul
child-NOM.HON letter-ACC
ssu-ess-ta.
write-PST-DECL

d. *Ai-kkeyse pyenci-lul
child-NOM.HON letter-ACC
ssu-si-ess-ta.
write-HON-PST-DECL

'The child wrote a letter.'

Examples in (3) show a full paradigm of subject honorifics with an honorable subject *kyoswunim* 'professor'. Interestingly, native speakers found these sentences acceptable even in the absence of honorific markings (Song et al., 2019).

(3) a. Kyoswunim-i pyenci-lul
professor-NOM letter-ACC
ssu-ess-ta.
write-PST-DECL

b. Kyoswunim-i pyenci-lul
professor-NOM.HON letter-ACC
ssu-si-ess-ta.
write-HON-PST-DECL

c. ?Kyoswunim-kkeyse pyenci-lul
professor-NOM.HON letter-ACC
ssu-ess-ta.
write-PST-DECL

d. Kyoswunim-kkeyse pyenci-lul
professor-NOM.HON letter-ACC
ssu-si-ess-ta.
write-HON-PST-DECL

'The professor wrote a letter.'

(3d) illustrates the canonical co-occurrence of both honorific markings *-kkeyse* and *-si*, representing the most expected honorific configuration. However, as shown in (3a), the sentence with an honorable subject without any honorific markings is considered acceptable to native speakers of Korean. Even more intriguingly, sentences are judged acceptable when the verbal honorific marker *-si* is present without *-kkeyse*, as in (3b). The reverse configuration, in which *-kkeyse* appears without *-si* as in (3c), is the least favorable but still considered acceptable, suggesting that native speakers exhibit considerable flexibility in honorific marking when the subject referent is honorable. The marginal acceptability is indicated by ? in the examples.

2.3 Honorifics and language models

Honorifics in language models have not yet been widely studied. Existing work has primarily focused on how the inclusion of honorifics affects model performance on downstream tasks, including text similarity task (Hwang et al., 2024), request speech act accuracy and friendliness (Jung et al., 2024), and pragmatic competence in request expressions (Chen et al., 2025).

This focus on task performance, however, leaves open the question of how models process the honorific system itself. To our knowledge, only two

Condition	Sentence					
a. HON-HON	Ecey	cenyek-ey	honcase	**kyoswunim**-i	pyenci-lul	ssu-**si**-ess-ta.
	yesterday	evening-LOC	alone	**professor**-NOM	letter-ACC	write-**HON**-PST-DECL
	'The professor wrote a letter alone yesterday evening.'					
b. HON-PLAIN	Ecey	cenyek-ey	honcase	**kyoswunim**-i	pyenci-lul	ssu-ess-ta.
	yesterday	evening-LOC	alone	**professor**-NOM	letter-ACC	write-PST-DECL
	'The professor wrote a letter alone yesterday evening.'					
c. PLAIN-PLAIN	Ecey	cenyek-ey	honcase	**ai**-ka	pyenci-lul	ssu-ess-ta.
	yesterday	evening-LOC	alone	**child**-NOM	letter-ACC	write-PST-DECL
	'The child wrote a letter alone yesterday evening.'					
d. PLAIN-HON	*Ecey	cenyek-ey	honcase	**ai**-ka	pyenci-lul	ssu-**si**-ess-ta.
	yesterday	evening-LOC	alone	**child**-NOM	letter-ACC	write-**HON**-PST-DECL
	'The child wrote a letter alone yesterday evening.'					

Table 1: Examples of Experiment 1 main conditions in CLOSE distance (where * indicates unacceptability)

studies have specifically investigated the Korean honorific system in language models from a linguistic perspective. On the one hand, Lee and Wang (2023) show that language models such as GPT-2, RoBERTa, BERT, and ALBERT fail to fully capture addressee honorifics in Korean. On the other hand, Noh et al. (2024) examine whether four Korean BERT-based models including KR-BERT and KoELECTRA process honorific mismatches in a human-like manner. They specifically focus on the presence of the verbal honorific marker *-si*, comparing the patterns in (2a)-(2b) against their honorable counterparts in (3a)-(3b). Their results demonstrate that Korean BERT models exhibit human-like sensitivity to honorific mismatches, suggesting that language models are capable of capturing certain morphosyntactic properties of the honorific system.

Our study investigates whether language models process subject honorifics in a human-like manner, building on Noh et al. (2024). We extend their work in two important aspects. First, whereas their study focused primarily on the presence of the verbal honorific marker *-si*, we employ a full dataset paradigm that also includes the nominative honorific marker *-kkeyse*, allowing for a more comprehensive and complete evaluation of subject honorifics. Second, while their analysis was limited to Korean-trained encoder models, we additionally examine multilingual and decoder-based models, thereby broadening the empirical coverage and enabling a more systematic assessment of how different model architectures and training regimes handle Korean honorifics.

3 Methods

We conducted three experiments to test whether language models process Korean subject honorifics similarly to humans.[2] The experiments build progressively: Experiment 1 establishes a baseline by replicating the experimental design of Noh et al. (2024), examining how verbal honorific marking *-si* interacts with subject honorability, without considering *-kkeyse*. Experiments 2 and 3 then examine the full subject honorifics paradigm by including both *-kkeyse* and *-si*. Importantly, Experiment 2 tests honorific patterns with honorable subjects, while Experiment 3 uses plain subjects. These design choices allow us to distinguish whether models are sensitive to surface dependency alone (i.e., co-occurrence of *-kkeyse* and *-si*), or whether they additionally encode the pragmatic constraints (i.e., honorability of the referent) that trigger subject honorifics.

3.1 Experiment 1

3.1.1 Stimuli

We manipulated whether the subject was honorable or plain, and whether the verb carried honorific marking *-si*. This created the four experimental conditions detailed in Table 1: HON-HON (honorable subject with honorific verb), HON-PLAIN (honorable subject with plain verb), PLAIN-PLAIN (plain subject with plain verb), and PLAIN-HON (plain subject with honorific verb). We used 10 honorable subjects, 10 plain subjects, and 10 verb pairs, each appearing with and without honorific marking *-si*. In addition, the distance between the subject and verb was manipulated by inserting a three-word adverbial phrase (e.g., *alone yesterday evening*), either between the subject and verb (FAR condition) or before the subject (CLOSE condition). This manipulation tested whether intervening material disrupts the processing of subject honorific

[2]The experimental stimuli and code used in this study are available at https://github.com/chloedkkwon/honorific-dependency.

Condition	Sentence
a. NO HONORIFICS	Kyoswunim-**i** ecey cenyek-ey honcase pyenci-lul ssu-ess-ta. professor-**NOM** yesterday evening-LOC alone letter-ACC write-PST-DECL 'The professor wrote a letter alone yesterday evening.'
b. VERB-ONLY	Kyoswunim-**i** ecey cenyek-ey honcase pyenci-lul ssu-**si**-ess-ta. professor-**NOM** yesterday evening-LOC alone letter-ACC write-**HON**-PST-DECL
c. SUBJECT-ONLY	?Kyoswunim-**kkeyse** ecey cenyek-ey honcase pyenci-lul ssu-ess-ta. professor-**NOM.HON** yesterday evening-LOC alone letter-ACC write-PST-DECL
d. ALL HONORIFICS	Kyoswunim-**kkeyse** ecey cenyek-ey honcase pyenci-lul ssu-**si**-ess-ta. professor-**NOM.HON** yesterday evening-LOC alone letter-ACC write-**HON**-PST-DECL 'The professor wrote a letter alone yesterday evening.'

Table 2: Examples of Experiment 2 main conditions in FAR distance (where ? indicates the least favorable option).

dependency patterns (e.g., Lee and Wang, 2023). The distance manipulation doubled the number of test sentences, yielding 800 sentences in total for Experiment 1 (4 conditions × 10 subjects × 10 verbs × 2 distances).

3.1.2 Measurement

We computed surprisal as a proxy for the acceptability of sentences, following proposals that it serves as a computational analog to processing difficulty in human reading (Hale, 2001; Levy, 2008). Under this framework, tokens that are less expected by the language model result in higher surprisal, reflecting the greater cognitive effort required when humans encounter unexpected input. In our study, surprisal was measured at the *verb* because it is the critical region for subject-verb honorific dependency. Thus, surprisal values were calculated by averaging across all tokens comprising the verb.

To calculate surprisal, we first obtained the probability distribution for each token position in the sentence. For masked language models, subword tokens of the target verb were masked individually to predict their probability. In causal language models, the distribution was based on all preceding tokens. Surprisal was then determined by taking the negative log probability of each subword token and averaging them across the entire verb.

$$Surprisal = -\log_2 P(\text{token}|\text{context}) \quad (1)$$

3.2 Experiments 2 and 3

3.2.1 Stimuli

Both Experiments 2 and 3 extended the design of Experiment 1 by adding *-kkeyse*, which occurs with honorable subjects in Korean. First, Experiment 2 used only honorable subjects, resulting in four conditions shown in Table 2: NO HONORIFICS (honorable subject without *-kkeyse* + plain verb), VERB-ONLY HONORIFICS (honorable subject without *-kkeyse* + honorific-marked verb), SUBJECT-ONLY HONORIFICS (honorable subject with *-kkeyse* + plain verb), and ALL HONORIFICS (honorable subject with honorific nominative marker *-kkeyse* + honorific-marked verb). We used the same 10 honorable subjects from Experiment 1 with or without *-kkeyse* and 10 verb pairs with or without the *-si* honorific marking.

Experiment 3 tested the same four conditions but with plain subjects (e.g., *ai* 'child'). In this case, NO HONORIFICS is the only acceptable condition. See Table 3 for example stimuli. We used 10 plain subjects with or without *-kkeyse* and the same 10 verb pairs. Both experiments included the same distance manipulation as Experiment 1, with a three-word adverbial phrase inserted either between the subject and verb (FAR condition) or before the subject (CLOSE condition). We generated 800 sentences (4 conditions × 10 subjects × 10 verbs × 2 distances) for each experiment.

3.2.2 Measurement

We calculated the surprisal of the verb as described in Experiment 1. In addition, as one of the reviewers pointed out, the mismatch between a plain subject with honorific marking in Experiment 3 may be most directly reflected in the subject region itself, rather than at the verb. To address this, we additionally calculated surprisal at the subject region (the subject noun and the nominative honorific marker *-kkeyse*) for Experiment 3, given that models consistently tokenized the subject noun and *-kkeyse* as a single unit.

For Experiment 2 and 3, we also calculated the accuracy, as the four conditions within each item set differ only in honorific marking, making surprisal most directly comparable within an item set.[3] Accuracy operationalizes this by measuring

[3]Note that accuracy is not calculated for Experiment 1, unlike Experiment 2 and 3, because the comparison between conditions is confounded by lexical differences between plain and honorable subjects (e.g., child vs. professor).

Condition	Sentence
a. ALL HONORIFICS	*Ai-**kkeyse** ecey cenyek-ey honcase pyenci-lul ssu-**si**-ess-ta. child-**NOM.HON** yesterday evening-LOC alone letter-ACC write-**HON**-PST-DECL
b. VERB-ONLY	*Ai-**ka** ecey cenyek-ey honcase pyenci-lul ssu-**si**-ess-ta. child-**NOM** yesterday evening-LOC alone letter-ACC write-**HON**-PST-DECL
c. SUBJECT-ONLY	*Ai-**kkeyse** ecey cenyek-ey honcase pyenci-lul ssu-ess-ta. child-**NOM.HON** yesterday evening-LOC alone letter-ACC write-PST-DECL
d. NO HONORIFICS	Ai-**ka** ecey cenyek-ey honcase pyenci-lul ssu-ess-ta. child-**NOM** yesterday evening-LOC alone letter-ACC write-PST-DECL 'The child wrote a letter alone yesterday evening.'

Table 3: Examples of Experiment 3 main conditions in FAR distance (where * indicates unacceptability).

whether the model's ranking of conditions within each item set aligns with human judgment. Thus, for Experiment 2, we awarded one point per item set if the model correctly assigned the highest surprisal to the SUBJECT-ONLY HONORIFICS condition, which is the least favorable option (marked with ? in Table 2). Accuracy was then computed as the number of points earned divided by the total number of item sets. For Experiment 3, a point was awarded if the models correctly predicted the lowest surprisal for the NONE HONORIFICS condition, which is the only acceptable option with plain subjects. Accuracy was calculated in the same way.

For each experiment, we fitted a linear mixed-effects model (LME) using *lme4* (Bates et al., 2015) in R (R Core Team, 2023), with surprisal at the target verb as the dependent variable and item set number as a random intercept. Model selection via AIC (Akaike, 1974) favored the maximal fixed-effects structure including all two- and three-way interactions between main condition, distance, and language model. We then used the *emmeans* package (Lenth, 2023) to examine differences within each factor of a model.

3.3 Language models

We tested seven language models: two encoder-based and five decoder-based. The encoder-based models were KoBERT (Kim, 2020) and KoELECTRA (Park, 2020), both based on BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) architectures, respectively, both with ~110M parameters. Although these models are smaller than the decoder-based models, no Korean encoder-based models with comparable (~1B) parameter counts were publicly available at the time of this study. Both models have been widely used in previous Korean NLP studies (Hur et al., 2021; Kim and Jo, 2022; Hwang and Oh, 2024), making them appropriate baselines.

The decoder-based models included three

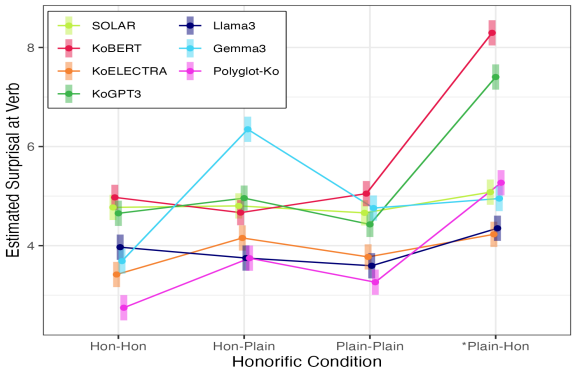


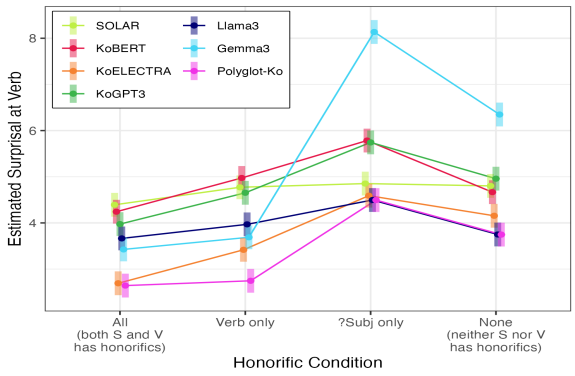
Figure 1: Estimated target verb surprisal across conditions in Experiment 1.

Korean-focused models: Polyglot-Ko (Ko et al., 2023) (1.3B parameters), trained exclusively on 863GB of Korean text; KoGPT 3 (SK Telecom, 2021) (1.2B), trained on a proprietary Korean corpus; and SOLAR (Kim et al., 2023) (10.7B), trained on a Korean- and English-focused multilingual corpus. We also included two multilingual models, Gemma 3 (Gemma Team, 2025) (1B) and Llama 3 (Grattafiori et al., 2024) (1B), trained on large multilingual datasets where Korean is present but not a primary focus. This allows us to compare Korean-focused models against multilingual models. Note that SOLAR is considerably larger than the others, as no smaller variant was publicly available.

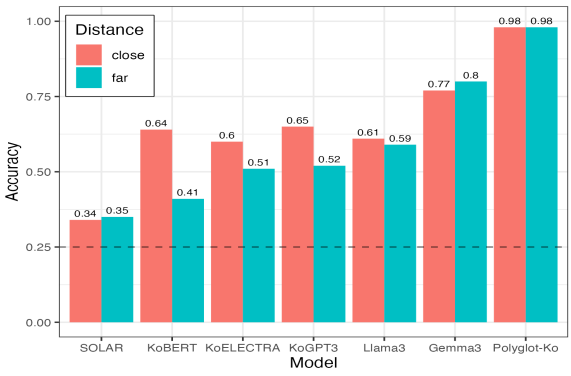
4 Results

4.1 Experiment 1

In Experiment 1, we compared surprisal across conditions that combine plain or honorable subjects and verbs with or without honorific marking *-si*. The results are presented in Figure 1, with colors representing each model. Most models assigned the highest surprisal to the unacceptable PLAIN-HON condition than to other conditions, except for Gemma 3 (light blue). This pattern is consistent



(a) Estimated target verb surprisal



(b) Model performance accuracy

Figure 2: Results of Experiment 2. In (b), the horizontal line indicates the chance level.

with Noh et al. (2024), although our results are different in that their HON-HON condition yielded higher surprisal than the HON-PLAIN condition, whereas we observed this pattern only in Gemma 3, and to a lesser extent, in KoELECTRA (orange). In this model, surprisals in the HON-PLAIN condition exhibited a wide distribution with a lower median than PLAIN-HON, rather than clearly elevated surprisal. In addition, the FAR condition showed slightly lower surprisal than the CLOSE condition for all models (mean difference: 0.49).

4.2 Experiment 2

In Experiment 2, we used honorable subjects that occur with or without *-kkeyse* and verbs with or without *-si*. The results are shown in Figure 2a. All models except SOLAR (yellow green) exhibited the highest surprisal in the SUBJECT-ONLY condition, where an honorable subject occurs with the subject honorific marking *-kkeyse* but with a plain verb (i.e., without *-si*), making the sentence least favorable by human subjects (Song et al., 2019).

Figure 2b shows the accuracy of each model by the distance condition, ordered by the model's overall accuracy. It shows that all models achieve beyond chance level accuracy (0.25). Note that surprisal results in Figure 2a and accuracy results in Figure 2b may not necessarily align, because surprisal values are averaged across items within each condition while accuracy is calculated per item set. For instance, KoELECTRA (orange) shows a left-to-right increase in surprisal consistent with human judgments, yet this does not entail high accuracy, as condition-level averaging captures the overall trend across items rather than performance on each individual item.

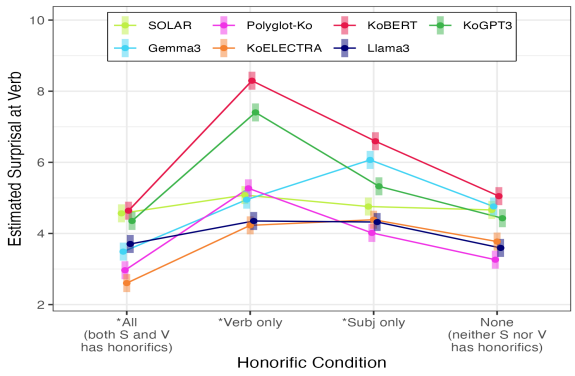
The Korean-focused decoder model Polyglot-Ko showed the highest accuracy, followed by the multilingual decoder model Gemma 3. Models' error patterns were distinctive: Polyglot-Ko nearly always predicted the highest surprisal for the SUBJ-ONLY condition, but its incorrect predictions (2%) consistently assigned the highest surprisal to the ALL condition, which is the most favorable option for human subjects. Gemma 3 exhibited a different pattern, usually assigning the highest surprisal to the NONE condition when wrong. These patterns contrast with all other models, which assigned the highest surprisal to the VERB-ONLY condition when wrong.

Regarding the distance condition, most models performed better in the CLOSE condition, likely reflecting sensitivity to subject-verb adjacency, while Polyglot-Ko and Gemma 3 showed similar accuracy across both conditions. Surprisal in the FAR condition was consistently lower than in the CLOSE condition for most models across all experiments: an intervening adverbial phrase between the subject and verb lowers surprisal across the board.

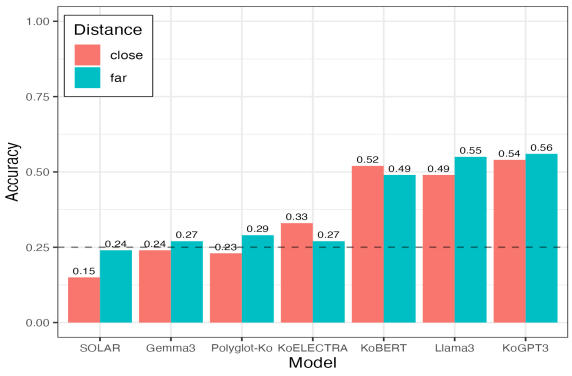
4.3 Experiment 3

The results from Experiment 3 demonstrate that most models generally exhibit high surprisal for the VERB-ONLY and SUBJECT-ONLY conditions at the verb region, as shown in Figure 3a. This aligns with human judgments that plain subjects such as *ai* 'child' are not expected to take honorific markings. However, the ALL condition, where both subject and verb are honorific-marked, yielded similar or even lower surprisal than the NONE condition in all models, despite the latter condition being the only acceptable option for native speakers.

Figure 3b presents model accuracy for Experiment 3. Again, a point was awarded if the mod-



(a) Estimated target verb surprisal



(b) Model performance accuracy

Figure 3: Results of Experiment 3 at the verb region. In (b), the horizontal line indicates chance level.

els predicted the lowest surprisal for the NONE condition since plain subjects are not expected to co-occur with any honorific marking. Overall accuracy was lower than in Experiment 2. More than half of the models achieve around or below the chance level accuracy (0.25). KoGPT 3 achieves the best performance by correctly predicting the lowest surprisal for the NONE condition in 55% of items across distance conditions. Unlike Experiment 2, error patterns were consistent: all models preferentially predicted the ALL condition as having the lowest surprisal. Note that the effect of distance varied by model type: encoder-based models (KoELECTRA and KoBERT) performed better in the CLOSE condition while decoder-based models performed better in the FAR condition, to varying degrees.

Figure 4 shows surprisals of each model at the subject region. KoELECTRA and decoder-based models, except for SOLAR, exhibited higher surprisals when plain subjects occurred with the honorific nominative marker, aligned with human judgments. Specifically, the ALL condition yielded higher surprisal than VERB-ONLY, and the SUBJECT-ONLY condition yielded higher surprisal than NONE condition. In contrast, KoBERT (red) showed the opposite pattern when comparing the ALL and VERB-ONLY conditions, yielding higher surprisal for VERB-ONLY than for ALL.

5 Discussion

This study examined whether seven language models have acquired knowledge of Korean subject honorifics, using surprisal as a proxy for acceptability across three experiments. The results reveal a systematic dissociation: models were generally good at detecting surface morphosyntactic

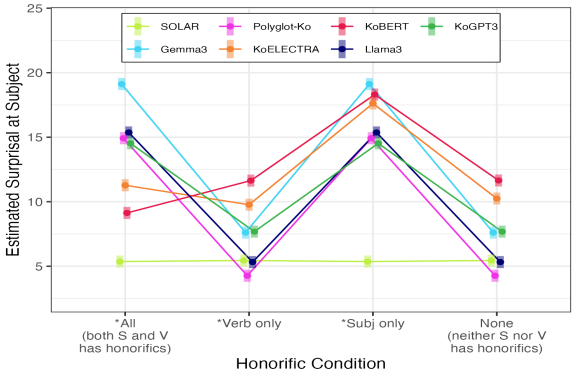


Figure 4: Estimated subject surprisal across conditions in Experiment 3.

mismatches but failed to respect the pragmatic constraints that govern honorific use. We discuss these findings in terms of what they reveal about the nature of linguistic knowledge in language models and broader implications for our understanding of language models' capabilities and limitations.

5.1 Surface heuristics vs. socio-pragmatic grounding

Experiments 1 and 2 demonstrated that most models assigned higher surprisal to less favorable honorific constructions (i.e., the PLAIN-HON condition in Experiment 1 and the SUBJECT-ONLY condition in Experiment 2). However, Experiment 3 revealed critical limitations. Most models, except SOLAR and KoBERT, correctly assigned higher surprisal to the subject region when plain subjects occurred with the honorific nominative marker, suggesting some sensitivity to the mismatch between the status of the subject and honorific marking at the subject position. However, models except Llama 3 incorrectly assigned the lowest surprisal to the verb

region when plain subjects occurred with both *-kkeyse* and *-si* (i.e., ALL condition), despite this condition being unacceptable to native speakers. Rather than treating the NONE condition as the most expected, as human speakers would, models appeared to generalize a surface-level preference for co-occurrence between subject nominative marker *-kkeyse* and verb honorific marker *-si*, irrespective of whether the subject referent warrants honorific treatment. These findings suggest that models have learned a surface co-occurrence heuristic between subject and verb honorific markers, instead of internalizing the underlying socio-pragmatic principle that honorific marking is licensed by the social status of the referent. In other words, models appear to prioritize surface distributional patterns over pragmatic appropriateness. Our results indicate that even Korean-focused models, which have extensive exposure to Korean honorific patterns, have not fully acquired this socio-pragmatic dimension.

These findings align with broader arguments in the literature that language models excel at learning distributional patterns but struggle with context-dependent reasoning that requires grounding in social knowledge (e.g., Sap et al., 2022). Korean honorifics present a particularly interesting test case for this distinction because the acceptability of a construction depends not just on morphosyntactic well-formedness but on a generalized understanding of the socio-pragmatic conditions under which honorific marking is licensed. This can be further illustrated by two types of cases that are difficult to account for on the basis of distributional patterns alone. First, honorific marking can occasionally be used with non-honorable referents such as pets, when the speaker playfully elevates the social status her pet (e.g., *Wuli koyangi-**kkeyse** pappu-**si**-ta*. roughly tranlated as 'His majesty, my cat, is busy.'). Second, even with an inherently honorable referent, a speaker may opt out of honorific marking entirely when she perceives herself as close and familiar with that referent (see also Song et al., 2019). Future work should construct experiments that directly manipulate such factors, which would allow us to probe whether language models can capture deeper level of pragmatic knowledge.

5.2 Distance sensitivity

A secondary finding across all three experiments was that the distance between subject and verb modulated model performance. Encoder-based models (KoBERT and KoELECTRA) consistently performed better in the CLOSE condition across all three experiments, while decoder-based models showed more variable patterns.

Interestingly, models generally showed lower surprisal in the FAR condition despite lower accuracy, presenting a dissociation between these two measures. One possibility is that lower surprisal reflects reduced sensitivity to the honorific dependency between the subject and verb. For instance, because intervening material dilutes the local context, the models may end up assigning less extreme surprisal values across all conditions, thereby reducing their ability to find unacceptable constructions more surprising. However, we could not conclude from the current data that the two patterns are directly related.

5.3 Korean-focused vs. multilingual models

Contrary to our hypothesis that Korean-focused models would consistently outperform multilingual models, the results did not show a clear advantage for Korean-focused training. For example, Llama 3, a multilingual model, achieved the second-highest accuracy in Experiment 3 and performed comparably to Korean-focused models. Meanwhile, SOLAR, despite being both Korean-focused and considerably larger than all other models, did not consistently outperform either multilingual or smaller models. These findings echo broader observations in the multilingual NLP literature that the relationship between language-specific training data and downstream linguistic performance is not straightforwardly linear (Conneau et al., 2020).

5.4 Limitations

One limitation of the present study is the absence of human judgment data for Experiment 3. Although it is a well-established fact in the literature that plain subjects do not trigger any honorific markings (e.g., Kim and Sells, 2007; Choi and Harley, 2019), empirical validation would strengthen the conclusions.

References

Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Yijun Chen, Peng Yue, and Henry Davidge. 2025. Comparing generative AI with native speakers in terms of request expressions in Japanese. *Discover Education*, 4(478).

Jaehoon Choi and Heidi Harley. 2019. Locality domains and morphological rules: Phases, heads, node-sprouting and suppletion in Korean honorification. *Natural Language & Linguistic Theory*, 37(4):1319–1365.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2021. Assessing BERT's ability to learn Italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, 14.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166. Association for Computational Linguistics.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

Yuna Hur, Suhyune Son Son, Midan Shim, Jungwoo Lim, and Heuiseok Lim. 2021. K-EPIC: Entity-perceived context representation in Korean relation extraction. *Applied Sciences*, 11(23):11472.

Sunik Hwang and Hayoung Oh. 2024. Trust beyond numbers: Data augmentation formula for poll prediction. *KSII Transactions on Internet and Information Systems*.

Yerin Hwang, Yongil Kim, Jeesoo Bang, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. Kosmic: Korean text similarity metric reflecting honorific distinctions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9954–9960, Torino, Italia. ELRA and ICCL.

Gayeon Jung, Joeun Kang, Fei Li, and Hansaem Kim. 2024. Are large language models affected by politeness? Focusing on request speech acts in Korean. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation (PACLIC 38)*, pages 1–10.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Jung, and 1 others. 2023. SOLAR 10.7b: Scaling large language models with simple yet effective depth up-scaling. *Preprint*, arXiv:2312.15166.

Gyeongmin Kim and Jaechoon Jo. 2022. Verification of a dataset for Korean machine reading comprehension with numerical discrete reasoning over paragraphs. *JOIV: International Journal on Informatics Visualization*, 6(2-2):587–592.

Jong-Bok Kim and Peter Sells. 2007. Korean honorification: a kind of expressive meaning. *Journal of East Asian Linguistics*, 16(4):303–336.

Sungjoon Kim. 2020. KoBERT. GitHub repository.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, jiwung Hyun, and Sungho Park. 2023. A technical report for Polyglot-Ko: Open-source large-scale Korean language models. *Preprint*, arXiv:2306.02254.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? A case study of reflexive anaphor licensing in English control constructions. *Proceedings of the Society for Computation in Linguistics*, 5(1):206–211.

Soo-Hwan Lee and Shaonan Wang. 2023. Do language models know how to be polite? *Society for Computation in Linguistics*, 6(1):375–378.

Russell V. Lenth. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.9.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kangsan Noh, Sanghoun Song, and Eunjeong Oh. 2024. How language models understand honorific mismatches in korean. *Language Research*, 60(3):303–322.

Jangwon Park. 2020. KoELECTRA: Pretrained ELECTRA model for Korean. GitHub repository.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

SK Telecom. 2021. Ko-GPT-Trinity 1.2b. Hugging Face model repository.

Sanghoun Song, Jae-Woong Choe, and Eunjeong Oh. 2019. An empirical study of honorific mismatches in Korean. *Language Sciences*, 75:47–71.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

A Abbreviations

ACC = Accusative case; DECL = Declarative sentential ending; HON = Verbal honorific morpheme; LOC = Locative postposition; NOM = Nominative case; NOM.HON = Honorific nominative case; PST = Past tense.