How Language Models Understand Honorific Mismatches in Korean

Kangsan Noh[1], SanghounSong[1†] & Eunjeong Oh[2†]
[1]Korea University, [2]SangmyungUniversity

ABSTRACT

This study investigates whether language models can process honorific mismatches in Korean, which occur when syntactic agreement in honorification is violated. Two types of mismatches are examined: YN, in which an honorific referent is paired with a non-honorific verb; and NY, in which a non-honorific referent is paired with an honorific verb. Previous studies showed that native speakers consider YN mismatches relatively acceptable but not NY mismatches. To understand the manner by which language models manage such patterns, surprisal-a complexity metric reflecting sentence likelihood-is applied to four Korean models: KR-BERT, KoELECTRA-base, KLUE-RoBERTa-base, and KLUE-RoBERTa-large. A dataset of 3,200 sentences is used to estimate surprisal for NN matches, NY mismatches, YN mismatches, and YY matches. The results show that the models primarily reflect human judgments, i.e., YN mismatches are considered acceptable, whereas NY mismatches are not. However, the models deviated from human-like processing in managing YY matches, where no violations occurred, likely because of the rarity of YY constructions in the training data. This suggests that, whereas the models demonstrate partial success in processing honorifics, they depend on statistical patterns and lack the deeper pragmatic understanding required for full syntactic and contextual competence.

Keywords: functional linguistic competence, honorification, language model, mismatch, surprisal

1. Introduction

The present study aims to see how language models process honorific mismatches in Korean. The Korean honorific system requires a conjugation between an honorable entity and the honorable marker *-si-*, as in (1).

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A01080365).
† Corresponding authors: sanghoun@korea.ac.kr, eoh@smu.ac.kr

(1) a. ai-ka o-(si)-ess-ta.
child-NOM come-(HON)-PST-DECL
'The child came.'
b. sensayng.nim-i o-(si)-ess-ta.
teacher(HON)-NOM come-(HON)-PST-DECL
'The teacher (honored) came.'

(Song et al., 2019: 53)

Honorific mismatches refer to instances where the required conjugation between an honorable entity and the honorable marker *-si-* is breached. For example, if the marker is applied, sentence (1a) becomes ungrammatical, as it involves a non-honorable entity (*the child*). In contrast, if the marker is absent, sentence (1b) becomes ungrammatical, as it involves an honorable entity (*the teacher*). In other words, honorific mismatches, caused by the presence or absence of the honorable marker *-si-*, affect the acceptability of sentences.

In this respect, honorification in Korean has been understood as analogous to subject-verb agreement in English and other European languages (Kang, 1988; Choe, 1988; Choi, 1993; Kim, 2012; Kim & Chung, 2015). However, there are also counterarguments against the consensus regarding the status of honorification as syntactic agreement (Namai, 2000; Choe, 2004; Ide, 2005; Kim & Sells, 2007). What is at issue is that honorific agreement in Korean and Japanese appears to be optional rather than compulsory, in contrast to English and other European languages (Boeckx & Niinuma, 2004). In this regard, Song et al. (2019) argue that honorification in Korean is a semantic/pragmatic phenomenon rather than a syntactic one, focusing on honorific mismatches in the language.

Recent progress in language models has led linguists to investigate how these artificial systems handle the complex hierarchical features of natural languages. While some theoretical linguists remain skeptical about the relevance of language models to the study of natural languages, their application goes beyond merely testing sophisticated artifacts (Linzen & Baroni, 2021). Language models are relevant to linguistics because, in the appropriate experimental context, these artificial learners can serve as valuable tools for testing existing linguistic hypotheses and refining linguistic theories about human linguistic competence.

Given the debates about the status of the honorific system in Korean, we aim to introduce a new empirical approach to honorification studies by using non-human

subjects: language models. Although subject-verb agreement in English and other European languages has been extensively studied (Linzen et al., 2016; Jawahar et al., 2019; Goldberg, 2019; Finlayson et al., 2021; Guarasci et al., 2023), the investigation of how language models handle agreement in Korean remains unexplored. To address this, we use a series of pretrained Korean language models to investigate how they handle honorific mismatches in Korean.

2. Background

2.1. Korean Honorific System

Honorification refers to the use of special linguistic forms to convey deference towards a referent or addressee. The Korean honorific system employs three methods for expressing honorifics: (pro)nouns, inflection, and suppletives (Song et al., 2019). For nouns, the honorific suffixes *nim* or *ssi* are attached if the referent is an honoree. For example, *sensayng* 'the teacher' is the plain form, while *sensayng.nim* 'the teacher (honored)' is the honorific form. Regarding inflection, honorific markers are used in both verbal and nominal forms. In verbal inflection, the marker *-si-* is added after the verbal stem, and in nominal inflection, the markers *-kkeyse* or *-kkey* are attached to the honoree. Lastly, suppletive verbs are used to indicate that the referent is regarded with respect by the speaker. For example, the verb *ca-* 'to sleep' is the neutral form. When referring to an honoree, its suppletive counterpart *cwumwusi-* is used instead.

2.2. Honorific Mismatches

Honorific mismatches refer to the cases where the syntactic agreement in honorification is violated. Specifically, there are two types of honorific mismatches: YN mismatches and NY mismatches. The former occurs when there is an honorific referent paired with a non-honorific verb, as in (2a). In contrast, the latter involves a non-honorific referent paired with an honorific verb, as in (2b).

(2) a. sensayng.nim-i o-ess-ta.
child-NOM come-PST-DECL
'The teacher came.'

b. ai-ka o-si-ess-ta.
child-NOM come-HON-PST-DECL
'The child came (honored).'

Concerning the two types of mismatches, Song et al. (2019) argue that there are logically four possibilities. The first possibility is that both YN and NY mismatches are unacceptable to Korean speakers. In this case, it can be concluded that honorification in Korean functions similarly to number agreement in English and other European languages. The second possibility is that both YN and NY mismatches are acceptable to Korean speakers, suggesting that Korean honorification is more of a stylistic condition rather than a strict syntactic agreement. The third possibility is that YN mismatches are unacceptable, while NY mismatches are generally acceptable. In this scenario, the previous syntactic argument for honorification remains valid. This is because previous syntactic studies advocating AGREE in honorification propose that the honorific referent triggers agreement, and disallowing YN mismatches supports this claim (Sakai & Ivana, 2009; Kim, 2012, 2017). The final possibility is the opposite of the third, where only NY mismatches are unacceptable. In this case, the syntactic argument for honorification would no longer hold, and Korean honorification would need to be explained from a different theoretical perspective, instead of a syntactic one.

With these four possibilities in mind, Song et al. (2019) conducted acceptability judgments with 382 native Korean speakers. The overall results indicated that the last possibility is empirically supported. Specifically, most native Korean speakers accepted YN mismatches in the acceptability judgments. Furthermore, corpus analysis showed that YN mismatches occurred twice as frequently as YY matches.

3. Design

The present study replicates the experiments conducted by Song et al. (2019). What sets this study apart is the use of neural language models instead of human subjects, along with the construction of test sentences. Our research questions are as follows: First, is the processing of honorification in Korean by language models similar to or different from that of native Korean speakers? Second, if language models process honorification differently from humans, how are these differences demonstrated?

3.1. Materials

Following the experimental design by Song et al. (2019), we created two types of datasets: one with subject referents (subject honorification) and another with object referents (object honorification). Subject honorification is categorized into regular and suppletive forms based on the verb type, while object honorification is divided into direct and indirect objects depending on the object type.

Table 1. Structure of the datasets

Type	Subtype	Sentences
Subject Honorification	Regular Forms	800
	Suppletive Forms	800
Object Honorification	Direct Objects	800
	Indirect Objects	800
Sum		3,200

In our test sentences, *-kkeyse* was not used. Song et al. (2019) reported that the impact of *-kkeyse* did not have a significant effect on acceptability judgments. Their analysis of the Sejong Spoken Corpus reveals that *-kkeyse* was rarely used, with only 13 instances found among 422,865 words. When it comes to neural language models, the statistical rarity of *-kkeyse* might distort the calculation of surprisal values, making it difficult to account for the effect of other factors. Specifically, we calculated the frequency of *-kkeyse* and other markers in the 2022 NIKL Korean Dialogue Corpus, distributed by the National Institute of Korean Language (NIKL, 2023). The results show that *-kkeyse* accounts for only 0.006% of the entire dataset (273 occurrences out of a total of 4,463,005 eojeols). Given the rarity of *-kkeyse*, we excluded it from our test sentences and focused on whether subjects/objects and verbs were matched in terms of honorification.[1)]

3.1.1. Subject Honorification: Regular Forms

1) One reviewer observed that the occurrences of *-kkeyse* might still be significant and recommended further corpus analysis, particularly since written corpus data was not examined. Although this is a valid critique, we exclude *-kkeyse* to maintain consistency with the experimental design of Song et al. (2019).

In regular forms, honorification is achieved by attaching the honorific marker *-si-* after the verbal stem. The basic structure of regular forms includes four conditions: NN, NY, YN, and YY. First, NN refers to cases where both subjects and verbs are non-honorific. Second, NY denotes cases where non-honorific subjects are paired with honorific verbs. Third, YN represents cases with honorific subjects and non-honorific verbs. Lastly, YY indicates cases where both subjects and verbs are honorific. Exemplary cases of each condition are shown in (3).

(3) a. ai-ka o-ess-ta.
child-NOM come-PST-DECL
'The child came.' (NN)
b. ai-ka o-si-ess-ta.
child-NOM come(HON)-PST-DECL
'The child came. (honored)' (NY)
c. sensayng.nim-i o-ess-ta.
teacher(HON)-NOM come-PST-DECL
'The teacher came.' (YN)
d. sensayng.nim-i o-si-ess-ta.
teacher(HON)-NOM come(HON)-PST-DECL
'The teacher came. (honored)' (YY)

3.1.2. Subject Honorification: Suppletive Forms

In suppletive forms, honorification is conveyed through the use of suppletive verbs. The basic structure of suppletive forms also includes four conditions: NN, NY, YN, and YY. Exemplary cases of each condition are shown in (4).

(4) a. ai-ka ppang-ul mek-nun-ta.
child-NOM bread-ACC eat-PRE-DECL
'The child eats bread.' (NN)
b. ai-ka ppang-ul capswusi-n-ta.
child-NOM bread-ACC eat(HON)-PRE-DECL
'The child eats (honored) bread.' (NY)
c. sensayng.nim-i ppang-ul mek-nun-ta.
teacher(HON)-NOM bread-ACC eat-PRE-DECL
'The teacher eats bread.' (YN)

d. sensayng.nim-i ppang-ul capswusi-n-ta.
teacher(HON)-NOM bread-ACC eat(HON)-PRE-DECL
'The teacher eats (honored) bread.' (YY)

3.1.3. Object Honorification: Direct Objects

When the honored referents are direct objects, the basic structure includes four conditions: NN, NY, YN, and YY. Exemplary cases of each condition are shown in (5).

(5) a. haksayng-i ai-lul manna-ss-ta.
student-NOM child-ACC meet-PST-DECL
'The student met the child.' (NN)
b. haksayng-i ai-lul poy-ess-ta.
student-NOM child-ACC meet(HON)-PST-DECL
'The student met (honored) the child.' (NY)
c. haksayng-i sensayng.nim-ul manna-ss-ta.
student-NOM teacher(HON)-ACC meet-PST-DECL
'The student met the teacher.' (YN)
d. haksayng-i sensayng.nim-ul poy-ess-ta.
student-NOM teacher(HON)-ACC meet(HON)-PST-DECL
'The student met (honored) the teacher.' (YY)

3.1.4. Object Honorification: Indirect Objects

When the honored referents are indirect objects, the basic structure also includes four conditions: NN, NY, YN, and YY. Exemplary cases of each condition are shown in (6).

(6) a. haksayng-i ai-eykey senmwul-ul cwu-ess-ta.
student-NOM child-DAT gift-ACC give-PST-DECL
'The student gave a gift to the child.' (NN)
b. haksayng-i ai-eykey senmwul-ul tuli-ess-ta.
student-NOM child-DAT gift-ACC give(HON)-PST-DECL
'The student gave (honored) a gift to the child.' (NY)

c. haksayng-i sensayng.nim-eykey senmwul-ul cwu-ess-ta.
student-NOM teacher(HON)-DAT gift-ACC give-PST-DECL
'The student gave a gift to the teacher.' (YN)

d. haksayng-I sensayng.nim-eykey senmwul-ul tuli-ess-ta.
student-NOM teacher(HON)-DAT gift-ACC give(HON)-PST-DECL
'The student gave (honored) a gift to the teacher.' (YY)

3.2. Language Models

We utilized encoder language models specifically designed for natural language understanding. In particular, we employed representative Korean language models built on the Bidirectional Encoder Representations from Transformers (BERT) architecture, a bidirectional version of transformer networks that accounts for both the left and right contexts of a masked word (Devlin et al., 2019). The models used include KR-BERT (Lee et al., 2020), KoELECTRA (Park, 2020), and KLUE-RoBERTa (Park et al., 2021). Specifically, we employed the base version of KoELECTRA and both the base and large versions of KLUE-RoBERTa. In sum, we utilized four distinct Korean language models: KR-BERT, KoELECTRA-base, KLUE-RoBERTa-base, and KLUE-RoBERTa-large.

3.3. Surprisal

Surprisal, which is the logarithm of the inverse of a probability (Tribus, 1961), indicates the informational value of a sentence. In computational linguistics, it is used as a metric to assess the complexity of processing a linguistic expression (Hale, 2001, 2016; Levy, 2008). For example, Futrell et al. (2019) define the surprisal $S(x_i)$ for a target word x_i as shown in (7).

(7) $S(x_i) = -\log_2(p(x_i | h_{i-1}))$

(Futrell et al., 2019: 33)

In (7), surprisal $S(x_i)$ is defined as the logarithm of the inverse probability of the word x_i, given h_{i-1}, which represents the language model's hidden state before the word x_i. Lower probabilities result in higher surprisal values, reflecting how improbable a sentence is according to the language model's probability distribution. In other words, sentences with lower probabilities receive higher surprisal values,

as they are less common in the corpus data and more likely to be ungrammatical.

Surprisal measurement typically targets a specific word in a sentence by masking its position. In this study, however, we account for the entire sentence by calculating the mean surprisal value for each word in the sentence. This approach allows us to compare all four conditions (NN, NY, YN, YY) simultaneously.

(8) a. ai-ka [MASK].
child-NOM [MASK]
'The child [MASK].'
b. sensayng.nim-i [MASK].
teacher(HON)-NOM [MASK]
'The teacher [MASK].'

Comparing the conditions NN and NY is possible by masking the position of verbs in (8a). Non-honorific verbs are inserted in [MASK] for the NN condition, while honorific verbs are inserted in [MASK] for the NY condition. Similarly, comparing the conditions YN and YY is possible by masking the position of verbs in (8b). Non-honorific verbs are inserted in [MASK] for the YN condition, while honorific verbs are inserted in [MASK] for the YY condition.

4. Results

The results for each model are provided in the following sections. As discussed earlier, Song et al. (2019) showed that Korean speakers found NY mismatches unacceptable, while YN mismatches were deemed acceptable. If the processing of honorification by neural language models resembles that of human subjects, the surprisal estimates for NY mismatches should be higher than those for YN mismatches. Therefore, we focused on comparing two types of honorific mismatches (NY vs. YN) to determine whether the models process honorific mismatches in a human-like manner.

4.1. KR-BERT

Figure 1 shows the distribution of surprisal values generated by KR-BERT. Table 2 presents the results of a paired t-test comparing the two types of mismatches: NY and YN. Cases where the models' responses were human-like are highlighted in boldface.

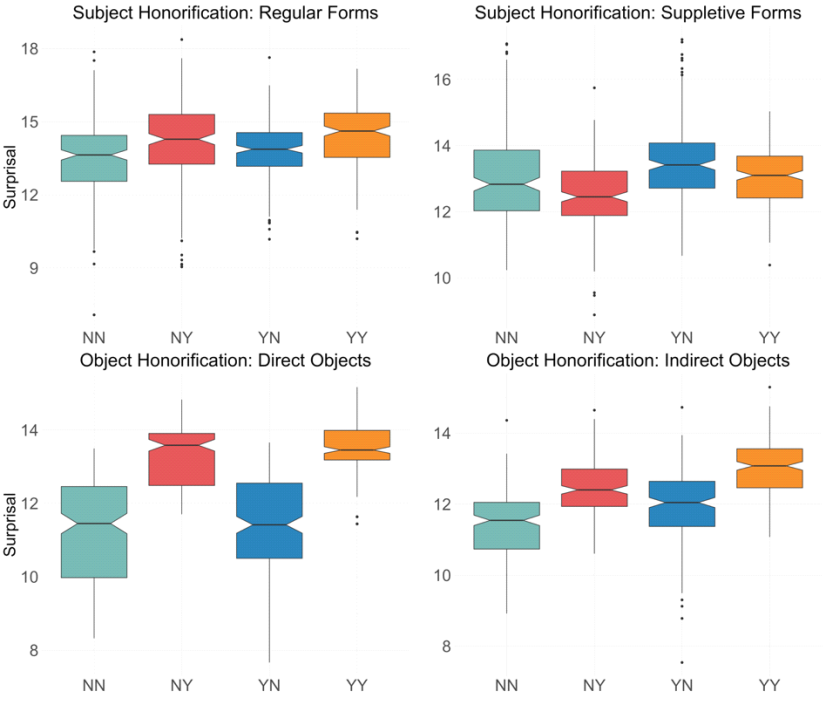

Figure 1. Surprisal distribution (KR-BERT)

Table 2. Paired t-test comparing NY and YN (KR-BERT)

Honorification Type		t statistic	p-value
Subject	Regular Forms	**2.5874**	**0.01038 (*)**
	Suppletive Forms	-10.49	< 0.001 (***)
Object	Direct Objects	**12.341**	**< 0.001 (***)**
	Indirect Objects	**7.4242**	**< 0.001 (***)**

Except for one condition (Subject Honorification: Suppletive Forms), the surprisal estimates for NY mismatches were higher than those for YN mismatches.

4.2. KoELECTRA-base

Figure 2 shows the distribution of surprisal values generated by KoELECTRA-base. Table 3 presents the results of a paired t-test comparing the two types of mismatches:

NY and YN. Cases where the models' responses were human-like are highlighted in boldface.

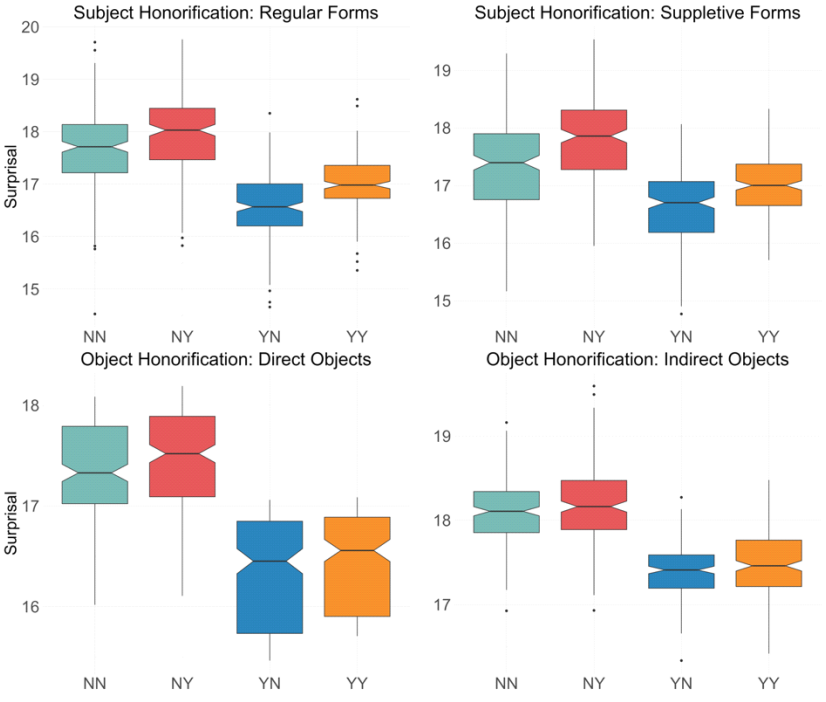

Figure 2. Surprisal distribution (KoELECTRA-base)

Table 3. Paired t-test comparing NY and YN (KoELECTRA-base)

Honorification Type		t statistic	p-value
Subject	Regular Forms	**31.383**	**< 0.001 (***)**
	Suppletive Forms	**23.694**	**< 0.001 (***)**
Object	Direct Objects	**31.251**	**< 0.001 (***)**
	Indirect Objects	**36.808**	**< 0.001 (***)**

The results reveal that the surprisal estimates for NY mismatches were higher than those for YN mismatches in all conditions.

4.3. KLUE-RoBERTa-base

Figure 3 shows the distribution of surprisal values generated by KLUE-RoBERTa-base. Table 4 presents the results of a paired t-test comparing the two types of mismatches: NY and YN. Cases where the models' responses were human-like are highlighted in boldface.

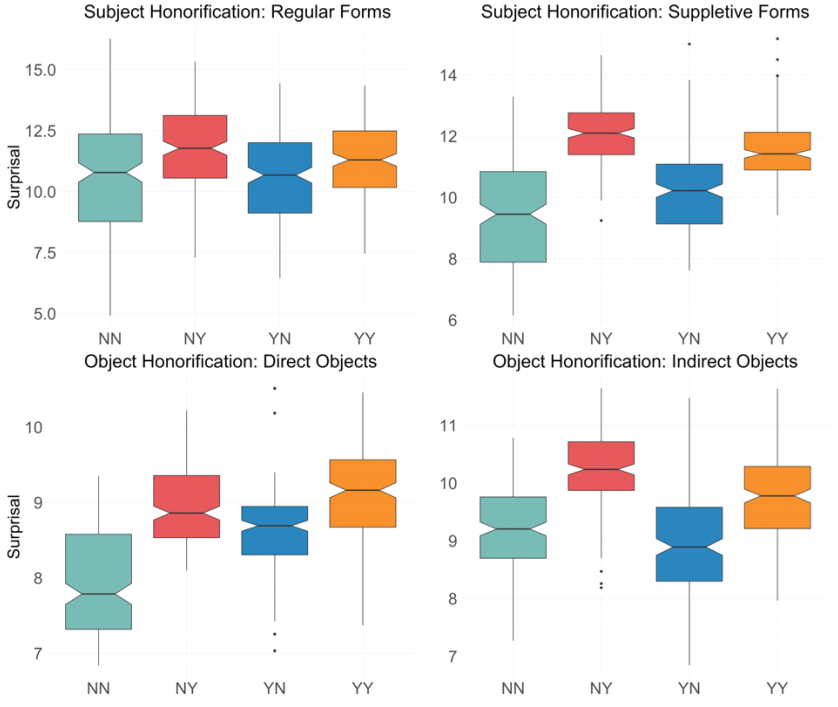

Figure 3. Surprisal distribution (KLUE-RoBERTa-base)

Table 4. Paired t-test comparing NY and YN (KLUE-RoBERTa-base)

Honorification Type		t statistic	p-value
Subject	Regular Forms	**13.933**	**< 0.001 (***)**
	Suppletive Forms	**22.352**	**< 0.001 (***)**
Object	Direct Objects	**5.3648**	**< 0.001 (***)**
	Indirect Objects	**18.511**	**< 0.001 (***)**

The results reveal that the surprisal estimates for NY mismatches were higher than those for YN mismatches in all conditions.

4.4. KLUE-RoBERTa-large

Figure 4 shows the distribution of surprisal values generated by KLUE-RoBERTa-large. Table 5 presents the results of a paired t-test comparing the two types of mismatches: NY and YN. Cases where the models' responses were human-like are highlighted in boldface.

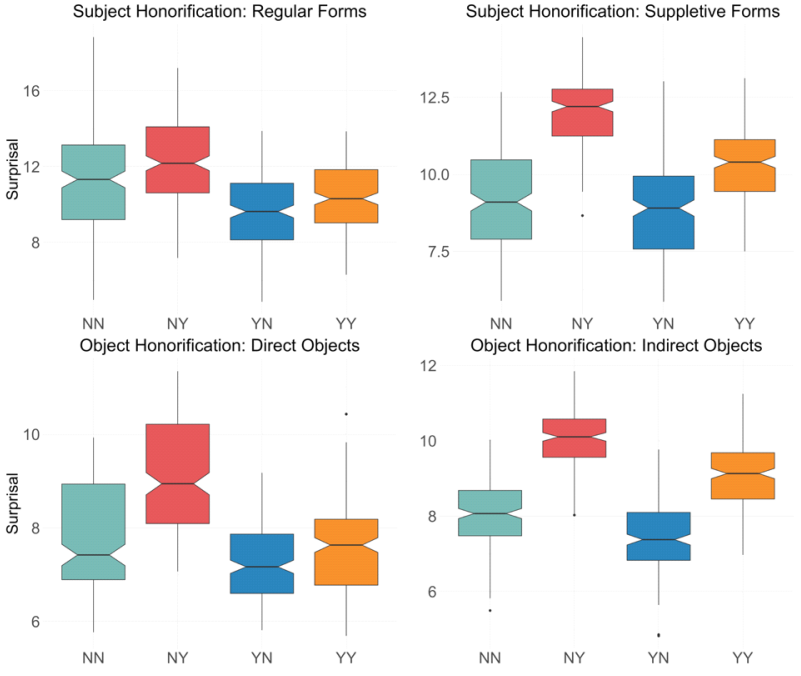

Figure 4. Surprisal distribution (KLUE-RoBERTa-large)

Table 5. Paired t-test comparing NY and YN (KLUE-RoBERTa-large)

Honorification Type		t statistic	p-value
Subject	Regular Forms	**25.023**	**< 0.001 (***)**
	Suppletive Forms	**35.062**	**< 0.001 (***)**
Object	Direct Objects	**19.225**	**< 0.001 (***)**
	Indirect Objects	**38.733**	**< 0.001 (***)**

The results reveal that the surprisal estimates for NY mismatches were higher than those for YN mismatches in all conditions.

5. Discussion

5.1. Where Language Models Are Human-like

In the comparison between the two types of honorific mismatches (NY vs. YN), all the neural language models behaved in a human-like manner. The models found NY mismatches to be unlikely but YN mismatches to be likely. Thus, the models' responses align with those of native Korean speakers; both human and model judgments support the last possibility that only NY mismatches are unacceptable. Consequently, the results of the present study provide additional empirical evidence against the syntactic argument for honorification.

5.2. Where Language Models Are Not Human-like

In contrast to the comparison between NY and YN mismatches, another comparison reveals that the models were not always human-like. The comparison between YN mismatches and YY matches shows that the surprisal estimates for YY matches were generally higher than those for YN mismatches. Table 6 presents the results of a paired t-test comparing the two. Cases where the models' responses were human-like are highlighted in boldface.

Table 6. Paired t-test comparing YN and YY

Models	Honorification Type		t statistic	p-value
KR-BERT	Subject	Regular Forms	-7.3286	< 0.001 (***)
		Suppletive Forms	**5.1978**	**< 0.001 (***)**
	Object	Direct Objects	-15.125	< 0.001 (***)
		Indirect Objects	-20.365	< 0.001 (***)
KoELECTRA-base	Subject	Regular Forms	-17.956	< 0.001 (***)
		Suppletive Forms	-10.084	< 0.001 (***)
	Object	Direct Objects	-14.162	< 0.001 (***)
		Indirect Objects	-9.2817	< 0.001 (***)

Table 6. Continued

Models	Honorification Type		t statistic	p-value
KLUE-RoBERTa-base	Subject	Regular Forms	-13.212	< 0.001 (***)
		Suppletive Forms	-19.103	< 0.001 (***)
	Object	Direct Objects	-25.364	< 0.001 (***)
		Indirect Objects	-24.232	< 0.001 (***)
KLUE-RoBERTa-large	Subject	Regular Forms	-12.925	< 0.001 (***)
		Suppletive Forms	-20.937	< 0.001 (***)
	Object	Direct Objects	-10.416	< 0.001 (***)
		Indirect Objects	-52.732	< 0.001 (***)

The results show that the surprisal estimates for YN mismatches were higher than those for YY matches on only one occasion (KR-BERT/Subject Honorification: Suppletive Forms). In other words, the models found YY matches more unlikely than YN mismatches. This contrasts with Korean speakers' acceptability judgments from Song et al. (2019), where YY matches were considered more acceptable than YN mismatches.

We argue that the contrast between human subjects and neural language models in the processing of YN mismatches and YY matches results from the statistical sparsity of YY matches. The corpus analysis reveals the statistical scarcity of YY matches. Song et al. (2019) extracted 950 utterances from the Sejong Spoken Corpus (https://ithub.korean.go.kr) where both the subject and verb are present, and one or both contain honorific markings. As a result, the 950 sentences were categorized into one of three types: YY matches, YN mismatches, or NY mismatches. The distribution of each type showed that YY matches (297 occurrences) were less frequent than YN mismatches (507 occurrences), while NY mismatches were the least frequent (146 occurrences).

In addition, the statistical rarity of suppletive verbs also may have played a role. For instance, we conducted a corpus analysis on the following three verbs: *mek-ta*, *tusi-ta*, and *capswu-si-ta*, all of which correspond to the English verb *eat*. While *mek-ta* is the neutral form, the latter two are suppletive forms. Since their past tense forms were used in our test material, we extracted the past tense forms of each verb from the 2022 NIKL Korean Dialogue Corpus (NIKL, 2023). Specifically, the verbal root and the past tense marker *-ess-* were taken into account during the extraction process

to capture all possible occurrences of each verb. The distribution of each verb is presented in Table 7.

Table 7. The distribution of *mek-ess*/*tusi-ess*/*capswusi-ess*

Verbs	*mek-ess*	*tusi-ess*	*capswusi-ess*
Frequency	2,473	6	0

The corpus analysis reveals that while *mek-ess*, the neutral form, occurred 2,473 times, its suppletive counterparts *tusi-ess* and *capswusi-ess* occurred 6 and 0 times, respectively. Thus, it is likely that the surprisal estimates for suppletive forms were influenced by their statistical rarity, leading to the models' judgment that YY matches are unacceptable.

5.3. Language Models' Cognitive Limitations

The results of the present study have shown that while the models are partially human-like in their processing of honorifics, their command of honorifics ultimately relies upon the statistical cues from the textual data. In both cases, whether the models behave in a human-like manner or not, statistical patterns have played a role. When the models behave like humans, previous corpus analysis revealed that YN mismatches (507 occurrences) are more frequent than NY mismatches (146 occurrences) in the corpus data (Song et al., 2019). Conversely, when the models are not human-like, the statistical rarity of YY matches seems to have influenced the models' responses. Thus, how the models processed honorification shows that their behavior is based on the statistical patterns of the textual data on which they were trained.

What matters here is that honorification in Korean is not simply a syntactic phenomenon, as highlighted by those who oppose the claim that it is a case of syntactic agreement (Namai, 2000; Choe, 2004; Ide, 2005; Kim & Sells, 2007; Song et al., 2019). Handling honorification requires a pragmatic understanding of the situations in which communication takes place. For instance, one should be able to decide whether to use honorifics with their counterpart. On one hand, a student may not use honorifics with his or her colleagues but may use them when communicating with seniors, who are usually older. On the other hand, the student may not use honorifics with their parents because of their close relationship, even though the parents are older.

Given the pragmatic nature of honorification, it is important to note that the models' formal linguistic competence does not necessarily imply functional linguistic competence. In this context, formal linguistic competence refers to the capacity for morphosyntactic knowledge (i.e., the ability to understand and apply the rules governing word structure and sentence formation), while functional linguistic competence refers to the capacity for social reasoning (i.e., the ability to navigate and comprehend social interactions), which is grounded in world knowledge. It has been suggested that these two competences do not always co-occur, even in the state-of-the-art language models (Mahowald et al., 2024). In other words, while LLMs exhibit remarkably strong formal linguistic competence, their performance on tasks involving functional linguistic competence continues to be inconsistent. We propose that the processing of honorifics may also fall under this distinction; although language models appear capable of processing morphosyntax, this does not imply they can also handle the subtle pragmatic reasoning required for honorification.

6. Conclusion

The purpose of this study was to examine how language models process honorific mismatches. The results revealed that the four Korean language models − KR-BERT, KoELECTRA-base, KLUE-RoBERTa-base, and KLUE-RoBERTa-large − exhibited human-like processing of YN and NY mismatches. YN mismatches were relatively acceptable, while NY mismatches were not. However, the results also showed that the models were not entirely human-like; they considered YY matches more unlikely than YN mismatches. This partial success in processing honorifics suggests that honorification is not merely a matter of syntactic agreement, but rather a pragmatic phenomenon requiring a nuanced understanding of the context and the agents involved.

Lastly, the limitations of this study need to be addressed. First, with recent trends in computational linguistics shifting from BERT-based encoder models to GPT-based decoder models, further analysis of decoder models' processing of honorifics appears to be a promising direction for future research. The primary reason this study focused on encoder models is that surprisal estimates can be calculated, which correspond to human acceptability judgments. We believe that further testing of decoder models can be conducted using a forced-choice test or Likert scale

assessment. Second, a more nuanced set of conditions for testing language models' processing of honorifics needs to be developed. This study used four types of conditions (NN matches, NY mismatches, YN mismatches, and YY matches) without providing additional contextual background, in order to compare with the results of native Korean speakers in Song et al. (2019). However, we believe that it would be possible to better assess language models' ability to handle honorifics by providing more detailed information about the agents and the communicative context in which the interactions occur. Taking these limitations and future research directions into account, we hope this study offers useful insights into both the linguistic capabilities of language models and the study of honorification.

References

Boeckx, C., & Niinuma, F. (2004). Conditions on agreement in Japanese. *Natural Language & Linguistic Theory*, 22 (3), 453–480.

Choe, H. S. (1988). Restructuring Parameters and Complex Predicates − A Transformational Approach. Massachusetts Institute of Technology. Ph. D. Dissertation.

Choe, J.-W. (2004). Obligatory honorification and the honorific feature. *Studies in Generative Grammar*, 14 (4), 545–559.

Choi, K. (1993). The structure of the long form negation construction in Korean. *Studies in Generative Grammar*, 3 (1), 25–78 [written in Korean].

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1 (Long and Short Papers), pp. 4171–4186), Minneapolis, Minnesota. Association for Computational Linguistics.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., & Belinkov, Y. (2021). Causal analysis of syntactic agreement mechanisms in neural language models. *arXiv preprint arXiv:2106.06087.*

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 32-42.

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

Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., & Esposito, M. (2023). Assessing BERT's ability to learn Italian syntax: A study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, *14*(1), 289-303.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397-412.

Ide, S. (2005). How and why honorifics can signify dignity and elegance: the indexicality of reflexivity of linguistic rituals. In: Lakoff, Robin T., Ide, Sachiko (Eds.), *Broadening the Horizon of Linguistic Politeness*. John Benjamin Publishing Company, Amsterdam, pp. 45–64.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Kang, M.-Y. (1988). *Topics in Korean syntax: Phrase structure, variable binding and movement*. Massachusetts Institute of Technology. Ph.D. Dissertation.

Kim, J., & Chung, I. (2015). A unified distributed morphology analysis of Korean honorification morphology. *Studies in Generative Grammar*, *25*(3), 631–650 [written in Korean].

Kim, J.-B., & Sells, P. (2007). Korean honorification: a kind of expressive meaning. *Journal of East Asian Linguistics*, *16*(4), 303–336.

Kim, Y.-H. (2012). Noun classes and subject honorification in Korean. *Linguistic Research*, *29*(3), 563–578.

Kim, Y.-H. (2017). The Korean honorific system and generative grammar: a reply to Kim and Chung (2015). *Studies in Modern Grammar*, *92*, 1-17 [written in Korean].

Lee, S., Jang, H., Baik, Y., Park, S., & Shin, H. (2020). KR-BERT: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979*.

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

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195-212.

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

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517-540.

Namai, K. (2000). Subject honorification in Japanese. *Linguistic Inquiry*, *31*(1), 170–176.

National Institute of Korean Language. (2023). NIKL Korean Dialogue Corpus (transcription) 2022(v.1.0). URL: https://kli.korean.go.kr/corpus

Park, J. (2020). KoELECTRA: Pretrained ELECTRA Model for Korean. https://github.com/monologg/KoELECTRA

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.-W., & Cho, K. (2021). KLUE: Korean Language Understanding Evaluation.

arXiv preprint arXiv:2105.09680.
Sakai, H., & Ivana, A. (2009). Rethinking functional parametrization: a view from honorification in the nominal domain in Japanese. *English Linguistics*, *26*(2), 437-459.
Song, S., Choe, J.-W., & Oh, E. (2019). An empirical study of honorific mismatches in Korean. *Language Sciences*, *75*, 47-71.
Tribus, M. (1961). Information theory as the basis for thermostatics and thermodynamics. *Journal of Applied Mechanics*, *28*(1), 1-8.

Kangsan Noh
Graduate Student
Department of Linguistics
Korea University
145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
E-mail: kasan1998@korea.ac.kr

Sanghoun Song
Associate Professor
Department of Linguistics
Korea University
145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea
E-mail: sanghoun@korea.ac.kr

Eunjeong Oh
Professor
Department of English Education
Sangmyung University
20 Hongjimun 2-gil, Jongno-gu, Seoul 03016, Korea
E-mail: eoh@smu.ac.kr

Received: October 30, 2024
Revised version received: December 2, 2024
Accepted: December 3, 2024