APPLING
International Applied Linguistics Research Conference

https://doi.org/10.29140/9781763711686-2

Rethinking LLM-Generated Texts as a New Linguistic Resource: The Case of Korean Politeness Expressions

Sungwon Dae
Independent Researcher, Republic of Korea

Abstract

This study investigates the potential of large language model (LLM)-generated texts as a new linguistic resource, particularly for language education, by examining the linguistic accuracy and sociopragmatic appropriateness of the usage of Korean politeness expressions, including honorifics and address terms. The main research question concerns what types of linguistic errors occur in LLM-generated texts in Korean, and what their implications are for the use of such texts as pedagogical resources. Specifically, this study evaluates the linguistic accuracy and sociopragmatic appropriateness of Korean politeness expressions in texts generated under LLM configurations. Korean politeness expressions encompass honorifics, speech levels, and terms of address, as their use is influenced by various factors, including interlocutors' age, social status, the degree of familiarity, and the communicative context. Korean politeness expressions are realized through various linguistic devices, including the honorific marker, particles, verb endings, honorific vocabulary, and terms of address. This study examines whether texts generated by four LLM configurations—Gemini 3.5 Flash, Gemini 3.1 Pro, GPT-5.5 Instant, and GPT-5.5 Thinking (High)—accurately and effectively reflect these linguistic devices. A set of 100 Korean-language prompts reflecting variations in social status, age, degree of familiarity, and communicative context was administered to each of the four LLM configurations, yielding 400 initial prompt submissions. As a result, of the 396 valid responses, 96 (24.24%) contained at least one error. Terms of address accounted for the largest proportion of coded error categories (35.71%), followed by inappropriate speech levels

E-mail: swdora109@gmail.com

In Y. Wang, J. Matthews, & B. Zou (Eds.), *Proceedings of the 2nd International Applied Linguistics Research Conference, 2026* (pp. 10–16). Castledown.

(21.43%). These findings indicate the need to review LLM-generated texts before using them as a new linguistic resource, such as instructional materials for language teaching.

Keywords: LLM-generated texts; Korean honorifics; politeness expressions; language education

Background

Recent advances in generative artificial intelligence have expanded the scope of research, including the use of large language model (LLM)-generated texts as linguistic resources in fields ranging from linguistic research to language education. For example, Bowden and Walker (2025) and Milička et al. (2026) proposed synthetic corpora for linguistic comparison and dialogue research. Lu and Zeng (2025) and Sekitani et al. (2025) explored the pedagogical potential of AI-generated texts and dialogues in language learning contexts. Regarding LLM-generated Korean texts, Jung and Kim (2026) found that Korean subject honorification patterns in the texts generated by ChatGPT were less varied than those observed in authentic spoken corpus data, raising questions about the extent to which LLM-generated texts reflect variation in actual language use. This study focuses on how accurately and effectively LLMs utilize Korean politeness expressions across four LLM configurations by evaluating their linguistic accuracy and sociopragmatic appropriateness for language education.

In Korean, the appropriate use of politeness expressions plays a crucial role in daily communication. The use of politeness expressions is shaped by various social and contextual factors, including age, social status, interpersonal relationships, and communicative situations (Brown, 2022). Korean politeness is realized through a combination of linguistic devices that often occur together in a single utterance: the honorific marker *-(eu)si-*, particles, verb endings, honorific vocabulary, and terms of address. Given the complexity of Korean politeness, LLM-generated texts should accurately reflect the linguistic forms and sociopragmatic conventions governing politeness usage to serve as reliable resources. Therefore, this study evaluates and analyzes the linguistic accuracy and sociopragmatic appropriateness of four different model configurations to assess their suitability as a linguistic resource.

Methodology Overview

This study evaluates the linguistic accuracy and sociopragmatic appropriateness of Korean politeness expressions in texts generated by four LLM configurations across various social contexts. The analysis focuses on three aspects of Korean politeness: honorific forms, speech levels, and terms of address. Based on the findings, this study discusses the potential limitations of using LLM-generated texts as linguistic resources for Korean language education.

This study analyzes texts generated by four model configurations: Gemini 3.5 Flash, Gemini 3.1 Pro, GPT-5.5 Instant, and GPT-5.5 Thinking (High). Data were collected between August 28, 2026, and August 31, 2026. All prompts were written in Korean and covered a wide range of text generation tasks, including role-play dialogues, emails, documents, messages, speech/presentations, and reflective writing. In addition, these prompts reflected diverse social contexts involving differences in age, social status, degree of familiarity, and communicative settings. Prompts were submitted to each model through the web interfaces of ChatGPT and Google Gemini. Each of the 100 prompts was initially submitted once to each of the four model configurations, but a response to one prompt was missing from one configuration. Therefore, to ensure comparability across models, the corresponding prompt was excluded from the analysis for all four configurations. Consequently, 99 responses per configuration were retained, yielding a total of 396 valid responses. For the evaluation and analysis, seven error categories related to Korean politeness expressions were selected: inappropriate speech levels, mixed speech levels, verb endings, honorific markers, particles, terms of address, and vocabulary. Errors unrelated to Korean politeness usage, such as general grammatical, spelling, and punctuation errors, were excluded from the politeness-related frequency analysis. Multiple error categories could be assigned to a single response.

Representative Results and Discussion

The same set of 100 prompts was applied to each of the four model configurations: Gemini 3.5 Flash, Gemini 3.1 Pro, GPT-5.5 Instant, and GPT-5.5 Thinking (High). Table 1 presents the frequency of politeness-related errors across the four configurations, categorized by error type.

The proportion of responses containing at least one politeness-related error was 23.23% (23/99) for Gemini 3.5 Flash, 19.19% (19/99) for Gemini 3.1 Pro, 27.27% (27/99) for GPT-5.5 Instant, and 27.27% (27/99) for GPT-5.5 Thinking (High). Among the four configurations, Gemini 3.1 Pro

showed the lowest proportion of error-containing responses, while both GPT-5.5 configurations showed the highest. Although GPT-5.5 Instant and GPT-5.5 Thinking (High) had identical proportions of error-containing responses, the latter had more error-category assignments (33 vs. 29). This suggests that similar proportions of error-containing responses do not necessarily correspond to the same number of coded error categories.

Table 1. Frequency of Politeness-Related Errors by LLM Configuration and Error Category

Category	Gemini 3.5 Flash	Gemini 3.1 Pro	GPT-5.5 Instant	GPT-5.5 Thinking (High)
Verb endings	1	2	6	2
Honorific marker	0	0	2	2
Particles	3	6	0	2
Vocabulary	3	1	3	4
Terms of address	10	6	10	14
Inappropriate speech levels	6	5	6	7
Mixed speech levels	3	4	2	2
Total coded errors	26	24	29	33
Responses with errors (n)	23	19	27	27
Valid responses (N)	99	99	99	99
Responses with errors (%)	23.23%	19.19%	27.27%	27.27%

Note. Frequencies represent politeness-related error-category assignments rather than individual error occurrences. Each error category was counted once per response, and multiple categories could be assigned to a single response. Therefore, the total number of category assignments may exceed the number of responses containing at least one politeness-related error. Percentages were calculated using 99 valid responses for each model configuration.

Table 2 presents the frequency and proportion of politeness-related error categories across the models. Terms of address constituted the most frequent politeness-related error category, accounting for 35.71%. This finding suggests that selecting contextually appropriate terms of address remains challenging for LLMs. Specifically, the errors in terms of address may derive from the

complexity of Korean address terms, in which appropriate forms vary depending on diverse factors, including age, gender, social status, and degree of familiarity. In families, kinship titles vary according to whether relatives belong to the maternal or paternal side of the family, as well as the marital status and gender of the interlocutors or the referent. In the workplace, terms of address are largely determined by social status and job rank. In school, terms of address differ according to age, gender, seniority, and degree of familiarity.

Table 2. Frequency and Proportion of Politeness-Related Error Categories Across Models

Rank	Error Category	Gemini	ChatGPT	Total	Proportion
1	Terms of address	16	24	40	35.71%
2	Inappropriate speech levels	11	13	24	21.43%
3	Mixed speech levels	7	4	11	9.82%
3	Particles	9	2	11	9.82%
3	Verb endings	3	8	11	9.82%
3	Vocabulary	4	7	11	9.82%
7	Honorific marker	0	4	4	3.57%
	Total	50	62	112	100%

Note. Frequencies represent politeness-related error-category assignments across the four model configurations (n = 112). Proportions were calculated by dividing the frequency of each category by the total number of category assignments. Because multiple error categories could be assigned to a single response, these proportions represent the distribution of coded error categories rather than the percentage of responses containing each error type. Tied categories were assigned the same rank.

The second most frequent error category was inappropriate speech levels, accounting for 21.43%, followed by mixed speech levels (9.82%). These results suggest that some responses contained inappropriate speech levels that could be considered overly polite or impolite, or exhibited inconsistencies in speech-level use.

Next, errors involving particles, verb endings, and vocabulary each accounted for 9.82%. Particle errors involved the use of non-honorific particles in contexts where honorific forms were expected to show respect to the subject, object, or referent. Verb-ending errors included the use of endings that did not align with the intended speech level of formality, such as using overly informal or formal endings. Vocabulary errors primarily involved the failure to use honorific vocabulary in contexts where it was expected.

Lastly, honorific-marker errors were the least frequent (3.57%), and these errors mostly included the omission of the honorific marker *-(eu)si-* in contexts where it was considered pragmatically expected.

These results show that LLMs face some challenges in generating linguistically accurate and appropriate Korean politeness expressions. This is likely to stem from the context-dependent nature of the Korean politeness system, which requires consideration of a variety of social and contextual factors and may consequently lead to inaccurate or unnatural outputs from LLMs. Therefore, LLM-generated texts should not be expected to consistently demonstrate linguistic accuracy and sociopragmatic appropriateness, and should be reviewed before they are utilized as a linguistic resource.

Conclusion and Limitations

This study analyzed errors in politeness expressions in LLM-generated texts. Among the 396 responses analyzed, 96 (24.24%) contained at least one error in the usage of Korean politeness expressions. Across the four LLM configurations, terms of address were the most frequently coded error category overall, followed by inappropriate speech levels. Errors involving particles, verb endings, vocabulary, and honorific markers were relatively infrequent but still occurred. These findings highlight the need for careful review before using LLM-generated texts as a linguistic resource, particularly for language education.

This study has several limitations. First, given that politeness expressions are not necessarily required in every context, these errors do not imply that LLM-generated texts are entirely inaccurate or inappropriate. In other words, not all linguistic devices need to be employed in every context to convey politeness. Song et al. (2019) demonstrated that certain honorific mismatches are acceptable through native-speaker acceptability judgments and spoken corpus analysis. Their research suggests that the usage of politeness expressions is influenced by semantic and pragmatic factors rather than being governed solely by syntactic agreement. Therefore, some expressions classified as errors in the present study may reflect acceptable linguistic variation depending on the context.

Second, the findings are limited to the specific model configurations evaluated at the time of data collection and may differ from those obtained using later versions. This is because LLM outputs may change over time as models are updated or replaced.

Last but not least, as each prompt was tested only once with each model configuration, the findings are also limited because the generated outputs may vary across repeated trials. Future research could incorporate repeated

prompt submissions and independent evaluations to assess the consistency of model outputs and strengthen the reliability of error analysis.

References

Bowden, K., & Walker, M. (2025). Dialogue scaffolding: Producing a realistic corpus of human-computer open-domain dialogues using a spoken dialogue system and ChatGPT. In F. Béchet, F. Lefèvre, N. Asher, S. Kim, & T. Merlin (Eds.), *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 538–560). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sigdial-1.44

Brown, L. (2022). Linguistic politeness. In A. S. Byon & D. O. Pyun (Eds.), *The Routledge handbook of Korean as a second language* (pp. 51–67). Routledge. https://doi.org/10.4324/9781003034704-5

Jung, Y., & Kim, K. M. (2026). A comparative analysis of AI-generated texts, corpus data, and speaker judgments: Subject honorification patterns in Korean. *Applied Corpus Linguistics, 6*(1), Article 100171. https://doi.org/10.1016/j.acorp.2025.100171

Lu, D., & Zeng, Y. (2025). Exploring the use of ChatGPT-generated model texts as a feedback instrument: EFL students' text quality and perceptions. *Innovation in Language Learning and Teaching*. Advance online publication. https://doi.org/10.1080/17501229.2025.2525341

Milička, J., Marklová, A., & Cvrček, V. (2026). AI Brown and AI Koditex: LLM-generated corpora comparable to human-written corpora of English and Czech texts. *Language Resources and Evaluation, 60*(4), Article 77. https://doi.org/10.1007/s10579-026-09939-3

Sekitani, K., Ogura, M., & Sato, T. (2025). Comparing ChatGPT and authentic discourse as L2 academic speaking partners: A corpus-based analysis and exploration of pedagogical applicability. *The EuroCALL Review, 32*(2), 75–87. https://doi.org/10.4995/eurocall.2025.23842

Song, S., Choe, J.-W., & Oh, E. (2019). An empirical study of honorific mismatches in Korean. *Language Sciences, 75*, 47–71. https://doi.org/10.1016/j.langsci.2019.101238