Beyond linearity and beyond English: Language model performance on Korean NPI licensing

Soo-Hwan Lee
Gyeongsang National University
soohwan.lee@gnu.ac.kr

Chanyoung Lee
Konkuk University
clee@konkuk.ac.kr

Gyu-Ho Shin
University of Illinois Chicago
ghshin@uic.edu

1 Introduction

Few studies have examined how large language models (LLMs) handle negative polarity items (NPIs) in languages other than English. This study investigates NPI licensing in Korean, a language that is subject to a well-established syntactic constraint. Our findings suggest that Korean-trained models exhibit clear difficulties in detecting this requirement when compared to human baseline judgments. This aligns with a persistent gap between human and model performance on syntax-sensitive phenomena in general (Wilcox et al. 2018; Gauthier et al. 2020; Hu et al. 2020; Warstadt et al. 2020; Lee and Schuster 2022). The present work reveals that NPI licensing is not adequately captured by the surprisal-based GPT language models: while the human results show a clear grammaticality contrast, the surprisal-based GPT models demonstrate poor and inconsistent sensitivity to this distinction. ChatGPT's sentence-level ratings align more closely with human judgments, although its performance leaves open the question of whether it has learned the underlying syntactic generalization. This study contributes to the growing body of literature on evaluating language models on linguistic dependencies and extending the evaluation to languages beyond English.

2 NPI licensing

NPIs depend on the presence of a licensor within their licensing domain. In Korean, negation expressed by *anh* 'not' typically licenses the NPI *amwuto* 'anyone' only when both elements occur within the same clause, a constraint known as the clausemate condition (Choe 1988; Kuno 1998). The clausemate condition is satisfied in (1) but violated in (2). In (1), both the NPI and its licensor (i.e., negation) appear in the embedded clause, satisfying the clausemate condition. In (2), the NPI appears in the embedded clause while its licensor occurs in the matrix clause, resulting in ungrammaticality (* = ungrammatical). Brackets indicate clause boundaries.[1]

(1) Yunho-ka [**amwuto** o-ci
Y.-NOM anyone come-NML
anh-ass-tako] sayngkakhay-ss-ta.
NEG-PST-COMP think-PST-DECL
'Yunho thought that no one came.'

(2) *Yunho-ka [**amwuto** wa-ss-tako]
Y.-NOM anyone come-PST-COMP
sayngkakha-ci **anh**-ass-ta.
think-NML NEG-PST-DECL
'Yunho did not think that anyone came.'

Korean and English differ with respect to how NPIs are licensed. In English, NPI licensing is natural when the NPI occupies an object position (3) but not when it occupies a subject position (4). Note that Korean is not subject to this restriction, as shown in (1).

(3) Maddie did **not** see **anyone**.

(4) ***Anyone** did **not** see Maddie.

This difference in where the NPI can appear offers a window into examining linguistic properties that cannot be tested in English. This highlights the importance of broadening the empirical scope of research to languages other than English.

3 Methods

3.1 Stimuli & design

We manipulated the structural positions of the NPI and its licensor in accordance with the clausemate condition, enabling identification of structural contexts in which licensing succeeds or fails across constructions. We designed four experimental conditions. In *npiM_negM*, both the NPI and its licensor appeared in the matrix clause, whereas in

[1] Abbreviations : NOM = nominative, COMP = complementizer, DECL = declarative, NEG = negation, NML = nominalizer, PST = past

NPI licensing	Condition	Example
✓	*npiE_negE*	윤호가 **아무도** 오지 **않**았다고 생각했다.
✗	*npiE_negM*	윤호가 **아무도** 왔다고 생각하지 **않**았다.
✗	*npiM_negE*	**아무도** 윤호가 오지 **않**았다고 생각했다.
✓	*npiM_negM*	**아무도** 윤호가 왔다고 생각하지 **않**았다.

Table 1: NPI licensing conditions and example stimuli.

npiE_negE, the two occurred in the embedded clause. In *npiM_negE*, the NPI was situated in the matrix clause and the licensor in the embedded clause. In *npiE_negM*, the NPI was positioned in the embedded clause and the licensor in the matrix clause. The first two conditions satisfy the clausemate requirement, whereas the latter two violate the requirement. Each condition contained 16 items, yielding a total of 64 sentences. Example stimuli that belong to the four conditions are provided in Table 1.

3.2 Human acceptability ratings

Fifty-six native Korean speakers (age: M = 23.5, SD = 2.6) participated in the experiment, which took approximately 30 minutes to complete. Participants completed an acceptability rating task on Qualtrics using a 6-point Likert scale, where 0 indicated 'very unacceptable' and 5 indicated 'very acceptable.' An excerpt of the instructions is provided below: 여러분께서는 주어지는 문장이 한국어에서 얼마나 자연스러운지를 6점 척도로 판단하시게 됩니다. 0점은 "매우 부자연스러움", 5점은 "매우 자연스러움"입니다. (English translation: 'Rate how natural the following sentence is in Korean on a six-point scale from 0 to 5, where 0 means 'very unnatural' and 5 means 'very natural.') Participants were instructed to respond as quickly as possible. Reaction times (RTs) were recorded to screen for inattentive responses.

3.3 Language models

3.3.1 Surprisal-based models

We employed three Korean-trained variants of the GPT architecture: KoGPT-2 (Korean GPT-2 model)[2], Ko-GPT-Trinity (Korean GPT-3 model)[3], and mGPT[4]. Table 2 provides detailed specifications of each model.

Our workflow involved loading three models from Hugging Face and processing sentences in plain-text format (one sentence per line). For each model, probabilities were computed conditional on the preceding context. Word-level surprisal was obtained by summing over constituent subwords, and sentence-level surprisal by aggregating across words, yielding fine-grained estimates of processing difficulty. Our workflow also standardized tokenization to ensure cross-architectural consistency. Primary adjustments included implementing left-padding—ensuring that models processed the full preceding context—and uniforming sentence boundary markers (<start> and <end> tokens) to maintain comparability across models.

	KoGPT-2	Ko-GPT-Trinity	mGPT
(i)	125M	1.2B	1.3B
(ii)	12	24	24
(iii)	12	16	16
(iv)	768	2,048	2,048
(v)	3,072	8,192	8,192

Table 2: Surprisal-based models. (i) = number of parameters, (ii) = number of layers, (iii) = number of heads, (iv) = hidden layer dimension, and (v) = FFN dimension.

To preserve the integrity of linguistic stimuli, sentences were segmented into eojeol via whitespace followed by subword tokenization. Eojeol refers to a Korean orthographic unit that contains one or more morphemes. It is often separated by spaces. Punctuation was retained to provide critical structural cues for clausal finality. Consequently, a sentence-final verb and its period (e.g., *mek-ess-ta.* / eat-PST-DECL.) were processed as a single unit, ensuring that surprisal metrics reflect naturalistic input.

Finally, models generated next-token probability distributions to calculate surprisal, defined as the negative logarithm of a token's conditional probability. Measured in bits, these values facilitate intuitive comparisons of information density. Sentence-level surprisal was derived by aggregating subword values across eojeols, serving as a computational proxy for linguistic acceptability: lower surprisal indicates higher perceived acceptability, whereas

[2]https://huggingface.co/skt/kogpt2-base-v2
[3]https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5
[4]https://huggingface.co/ai-forever/mGPT

higher surprisal indicates lower perceived acceptability.

3.3.2 Prompt-based models

We evaluated ChatGPT-5's sentence-level rating performance (conducted on 25 September 2025). The prompt was based on the instructions provided to our human participants. The actual prompt used in the experiment is as follows: 아래에 주어지는 문장이 한국어에서 얼마나 자연스러운지를 0부터 5까지 6점 척도로 판단해 줘. 0은 '매우 부자연스러움', 5는 '매우 자연스러움'이야. 정수(0, 1, 2, 3, 4, 5)로 판단을 하면서, 왜 그렇게 판단을 했는지 이유를 한 문장으로 말해줘. (English translation: 'Rate how natural the following sentence is in Korean on a six-point scale from 0 to 5. 0 means 'very unnatural' and 5 means 'very natural'. Use integers (0, 1, 2, 3, 4, 5) for your rating, and provide a one-sentence reason for your decision.')

For each item, the model produced a numerical rating along with a one sentence justification. Prompts were administered independently by two human annotators, and ChatGPT's memory was neither modified nor retained across trials.

4 Results

Figure 1 provides a summary of the overall results from the experiment.

4.1 Human rating

Table 3 presents the results of human ratings. A clear contrast emerged between grammatical and ungrammatical conditions: same-clause conditions received higher ratings than their cross-clause counterparts, with no reliable differences within either the grammatical or ungrammatical pairs.

Condition	Mean	SD	SE
npiE_negE	**4.224**	1.376	0.097
npiE_negM	**1.532**	1.500	0.105
npiM_negE	**1.608**	1.751	0.121
npiM_negM	**4.426**	1.069	0.075

Table 3: Human ratings by condition.

4.2 Surprisal-based models

Table 4 presents the results from the GPT language models. Notably, all three models assigned higher surprisal to the grammatical condition *npiM_negM* than to the ungrammatical condition *npiM_negE*, indicating a divergence between model behavior and human ratings.

Model	Condition	Mean	SD	SE
①	*npiE_negE*	**61.217**	4.244	1.061
	npiE_negM	**68.774**	4.595	1.149
	npiM_negE	**60.720**	5.016	1.254
	npiM_negM	**63.827**	3.909	0.977
②	*npiE_negE*	**79.192**	4.100	1.025
	npiE_negM	**86.482**	4.208	1.052
	npiM_negE	**81.182**	3.681	0.920
	npiM_negM	**82.409**	4.157	1.039
③	*npiE_negE*	**98.971**	5.213	1.303
	npiE_negM	**105.175**	5.791	1.448
	npiM_negE	**101.339**	4.882	1.220
	npiM_negM	**104.721**	5.163	1.291

Table 4: GPT model surprisal by condition. ① = KoGPT-2, ② = Ko-GPT-Trinity, and ③ = mGPT.

4.3 Prompt-based models

Table 5 presents the results of ChatGPT's acceptability ratings. Among the models examined, ChatGPT most closely matched human judgment patterns.

Condition	Mean	SD	SE
npiE_negE	**3.531**	0.562	0.140
npiE_negM	**2.375**	0.500	0.125
npiM_negE	**2.719**	0.315	0.079
npiM_negM	**4.063**	0.655	0.164

Table 5: ChatGPT ratings by condition.

5 Discussion

GPT surprisal values diverged from human ratings, revealing only partial and inconsistent encoding of the clausemate requirement. This pattern may reflect a reliance on surface-level distributional cues, such as frequency or linear adjacency between the NPI and negation, rather than the syntactic dependency established between the two. By contrast, ChatGPT aligned more closely with human judgments, assigning higher ratings to grammatical conditions than to ungrammatical conditions. This relative success may reflect the effects of scale, although future work is required to better examine whether the model captures the proper syntactic generalization. Future work also needs to explore the full range of NPIs and licensors in Korean to test whether models generalize beyond the specific lexical items used in this work. This will help us better determine whether LLMs are indeed sensi-

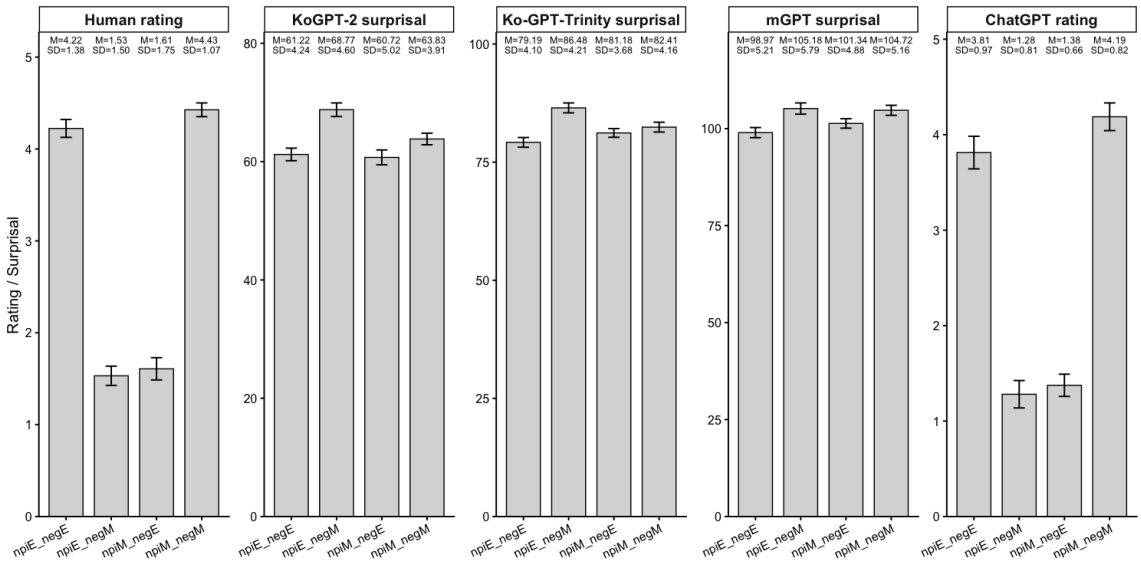

Figure 1: Human and language model results based on Tables 3–5.

tive to grammatical requirements rather than less abstract distributional cues.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful comments.

References

Hyon Sook Choe. 1988. *Restructuring Paramters and Complex Pedicates - A Transformational Approach*. Ph.D. thesis, Massachusetts Institute of Technology.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. 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.

Susumu Kuno. 1998. Negative polarity items in korean and english. *Cornell East Asia Series*, 98:87–132.

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. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 206–211, online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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.