GPT surprisal as a limited proxy for human sentence acceptability: Evidence from Korean clausal constructions

Soo-Hwan Lee (soohwan.lee@gnu.ac.kr)
Department of English Language and Literature, Gyeongsang National University
501 Jinju-daero, Jinju-si, Gyeongsangnam-do 52828, Republic of Korea

Chanyoung Lee (clee@konkuk.ac.kr)
Department of Korean Language and Literature, Konkuk University
120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea

Gyu-Ho Shin (ghshin@uic.edu)
Department of Linguistics, University of Illinois Chicago
601 South Morgan Street, Chicago, IL 60607, USA

Abstract

The present study compares human acceptability judgements of Korean clausal constructions—dative alternations, active–passive voice under animacy manipulation, and negative polarity item licensing—with surprisal estimates derived from GPT-style language models. Fifty-six native speakers evaluated 224 sentences, and surprisal was computed using three Korean-capable variants (KoGPT-2, Ko-GPT-Trinity, and mGPT). The models reproduced the strong preference for Dative–Accusative datives, captured voice- and animacy-related effects only partially, and failed to represent clause-bounded NPI licensing. These findings indicate limitations of surprisal as a proxy for acceptability beyond morphosyntax and motivate the development of cross-linguistic benchmarks for explainable AI and diagnostic evaluation.

Keywords: GPT, surprisal, human rating, clausal constructions, Korean

Introduction

A growing trend in the language sciences is the use of computational methods to investigate linguistic phenomena. This line of research examines the extent to which computational models simulate human language behavior (Chang, 2009; Hawkins et al., 2020; Jones & Bergen, 2024; Marvin & Linzen, 2019; Warstadt, Singh et al., 2019; Wilcox et al., 2018), while highlighting performance differences across algorithms and architectural variants (Contreras Kallens et al., 2023; Hu et al., 2020; Lee and Schuster, 2022; Lee, 2024; Shin & Mun, 2025; Warstadt & Bowman, 2020). Recent work further explores how model-based findings align with behavioral and corpus-based evidence that has illuminated core properties of human language architecture (Ambridge et al., 2020; Oh et al., 2022; Xu et al., 2023). However, the field remains heavily skewed towards English (Blasi et al., 2022; Chang & Bergen, 2024), a bias reinforced by the predominance of English-based Large Language Models (LLMs). This limits the generalizability of prior findings to underrepresented languages, constraining progress towards explainable AI—namely, evaluations that transparently relate model performance to human behavior.

The present study addresses this gap by examining whether Transformer-based models capture patterns of human sentence comprehension in Korean clausal constructions. Transformers are deep neural architectures that rely on self-attention to generate contextual representations and model long-range dependencies in parallel (Vaswani et al., 2017). When pretrained on large corpora (e.g., Generative Pre-trained Transformer, GPT; Radford et al., 2018), they produce robust probabilistic predictions. GPT, a decoder-only Transformer, embeds tokens with positional information and processes them through stacked attention and feed-forward layers with residual connections and normalization, conditioning predictions on prior context (Radford et al., 2018). Trained on web-scale data via next-token prediction, GPT functions as a general-purpose learner and exhibits zero- and few-shot generalization through prompting (Brown et al., 2020; Radford et al., 2019). GPT has become central to modeling human behavior: model probabilities and surprisal correlate with acceptability judgements and processing measures across tasks (Schrimpf et al., 2021; Warstadt & Bowman, 2020; Wilcox et al., 2018). Recent studies demonstrate GPT's grammatical sensitivity and metalinguistic awareness, with GPT-4 capturing effects of information structure on acceptability (Cuneo et al., 2025; Goldstein et al., 2022; Hosseini et al., 2022; Yoo et al., 2025).

Evaluation outcomes, however, vary across GPT families. GPT-2 closely tracks human preferences in clausal alternations and verb biases (Hawkins et al., 2020), whereas GPT-3 variants and ChatGPT yield mixed, prompt-sensitive results that limit their reliability for grammaticality assessment (Dentella et al., 2023; Hu & Levy, 2023; Lee & Wang, 2023). In reading-time prediction, surprisal generally correlates with processing difficulty via logarithmic cost–predictability relations; smaller GPT-2 models often outperform larger ones, which sometimes overfit (Shain et al., 2024; de Varda et al., 2024). Nonetheless, GPT-2 surprisal underestimates garden-path effects and embedded clause slowdowns, diverging from human patterns (Oh & Schuler, 2023). Large-scale benchmarks confirm these gaps: surprisal often underestimates syntactic ambiguity effects and accounts for limited item-level variance, suggesting that

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

next-word prediction alone does not fully capture processing demands (Huang et al., 2024). Broader metrics, such as sentence-level surprisal or discourse-level relevance, may better predict comprehension and reading speed across languages (Sun & Wang, 2025). At the representational level, probing studies show that LLM activations encode subtle theoretical distinctions, albeit with limited cross-lingual consistency, offering insight into internal mechanisms (Zhou et al., 2025). Overall, GPT models show qualified alignment with human sentence comprehension, with performance strongly shaped by model size and training design.

Korean is an agglutinative Subject–Object–Verb language characterized by overt case marking and flexible pre-verbal argument order (Sohn, 1999). As a typologically distant and underrepresented language, Korean provides a valuable test case for assessing the generalizability of prior findings. We draw on an open-access database of human acceptability ratings for clausal constructions (Shin et al., 2026), focusing on three construction types that probe the morphosyntax–semantics interface.

Target (1): Dative construction

Korean datives occur in two main variants with the canonical recipient-before-theme order: a Dative–Accusative pattern (1a), in which a dative-marked recipient precedes an accusative-marked theme, and an Accusative–Accusative pattern (1b), in which both recipient and theme bear accusative marking (Lee, 2014; Shin, 2016; Yoon, 2015).

(1) Korean dative construction: 'Mina gave a book to Jiho.'
a. Dative–Accusative
Mina-ka Jiho-eykey chak-ul cwu-ess-ta.
Mina-NOM Jiho-DAT book-ACC give-PST-DCL[1]
b. Accusative–Accusative
Mina-ka Jiho-lul chak-ul cwu-ess-ta.
Mina-NOM Jiho-ACC book-ACC give-PST-DCL

While both patterns encode transfer of possession (Haspelmath, 2015), studies reveal that the Accusative–Accusative pattern is rare and dispreferred compared to the Dative–Accusative form (Cho & Jeon, 2015; Kim & Shin, 2022). Because the accusative marker typically denotes themes (Choo & Kwak, 2008; Sohn, 1999), using it for recipients is often deemed less acceptable (Kim & Shin, 2022; Shin, 2020). Nevertheless, its occurrence in natural discourse confirms its grammaticality (Park & Yi, 2021). This alternation has been utilized to assess acceptability judgements (Cho & Jeon, 2015; Kim & Shin, 2022; Lee, 2014) and investigate cross-linguistic priming between Korean and English (Kim et al., 2020; Shin & Christianson, 2009), validating its status as a functional alternation pair.

Target (2): Passive construction

Passive voice, though cross-linguistically widespread (Haspelmath, 1990; Siewierska, 2013), is infrequent in Korean compared to the active (Park, 2021; Woo, 1997). The suffixal passive involves a nominative-marked theme subject, a dative-marked agent in an oblique position, and verbal morphology marking passivisation (Sohn, 1999). Active and passive constructions typically describe the same event but differ in perspectival prominence: actives foreground the agent, whereas passives promote the theme to subject and demote the agent to an oblique.

(2) Korean active vs. passive constructions
a. Human theme + Active voice
Kyenghuy-ka Cinho-lul cap-ass-ta.
Kyenghuy-NOM Cinho-ACC catch-PST-DCL
'Kyenghuy caught Cinho.'
b. Human theme + Passive voice
Cinho-ka Kyenghuy-eykey cap-hi-ess-ta.
Cinho-NOM Kyenghu-DAT catch-PSV-PST-DCL
'Cinho was caught by Kyenghuy.'
c. Inanimate theme + Active voice
Kyenghuy-ka kamca-lul cap-ass-ta.
Kyenghuy-NOM potato-ACC catch-PST-DCL
'Kyenghuy caught a potato.'
d. Inanimate theme + Passive voice
??Kamca-ka Kyenghuy-eykey cap-hi-ess-ta.
potato-NOM Kyenghuy-DAT catch-PSV-PST-DCL
'A potato was caught by Kyenghuy.'

Previous research shows that acceptability is modulated by animacy (Sohn, 1999; Yeon, 2003): actives readily allow both [+human] and [–human] themes, whereas passives degrade with [–human] themes. This asymmetry suggests that animacy interacts with argument structure and morphological marking to shape the perceived naturalness of Korean passives.

Target (3): Negative polarity item construction

Negation can license a negative polarity item (NPI), but only when the NPI and its licensor occur within the same licensing domain. This dependency is well documented in English and has attracted considerable attention in the NLP literature (Lee & Vu, 2024; Marvin & Linzen, 2018; Shin et al., 2023; Warstadt, Cao et al., 2019; Wilcox et al., 2019). In Korean, sentential negation *anh* 'not' typically licenses the NPI *amwuto* 'anyone' when both appear in the same clause, known as the clausemate condition (Choe, 1988; Kuno, 1998). This condition holds in (3a) but is violated in (3b), where the NPI occurs in the embedded clause and the licensor in the matrix clause, yielding ungrammaticality.

(3) Korean NPI construction
a. ✓clausemate condition
Mina-ka amwu-to o-ci
Mina-NOM anyone-FOC come-NML
ahn-ass-tako sayngkakhay-ss-ta.
NEG-PST-COMP think-PST-DCL

[1] ACC = accusative case marker; COMP = complementizer; DAT = dative marker; DCL = ending, declarative; FOC = focus marker; NEG = negation; NML = nominalizer; NOM = nominative case marker; PST = past tense marker; PSV = passive suffix.

'Mina thought that no one came.'

b. ✗clausemate condition

*Mina-ka amwu-to wa-ss-tako

Mina-NOM anyone-FOC come-PST-COMP

sayngkakha-ci anh-ass-ta.

think-NML NEG-PST-DCL

'Mina did not think that anyone came.'

Exploiting this dependency, we systematically vary the positions of licensors and licensees to examine how structural configuration affects NPI licensing. This aim is thus to identify the conditions under which licensing succeeds or fails across different compositions.

Methods

This study compares human acceptability ratings for these constructions with surprisal measures computed by GPT variants, treating sentence-level surprisal as a proxy for acceptability. The dataset was sourced from Shin et al. (2026): 96 datives (48 Dative–Accusative; 48 Accusative–Accusative), 64 passives (16 per condition, crossing theme animacy [human, inanimate] and voice [active, passive]), and 64 NPIs (16 per condition, crossing NPI location [matrix, embedded clause] and negation location [matrix, embedded clause]). All sentences used personal names and frequent, simple nouns as nominal arguments. Fifty-six native Korean speakers rated each sentence via Qualtrics on a six-point Likert scale (0 = "very unacceptable", 5 = "very acceptable"), responding as quickly as possible while maintaining accuracy. Reaction times, measured from sentence onset to scale selection, were used to identify inattentive responses.

GPT surprisal was computed via three open-access GPT variants available for Korean (Table 1): KoGPT-2[2], Ko-GPT-Trinity[3], and mGPT[4]. All modeling was conducted on a MacBook Pro (Apple M4 Max, 16-core CPU, 40-core GPU, 16-core Neural Engine, 128 GB unified memory).

Table 1. Model specifications

	KoGPT-2	Ko-GPT-Trinity	mGPT
Parameter #	125M	1.2B	1.3B
Layer #	12	24	24
Head #	12	16	16
Hidden layer dim	768	2,048	2,048
Feedforward network dim	3,072	8,192	8,192

Our workflow involved loading three models from Hugging Face, standardizing tokenizer behavior, and processing sentences in plain-text format (one sentence per line). Each sentence was segmented into *ejels* using whitespace, after which each ejel was tokenized into subword units by the model-specific tokenizer. Subword tokens were mapped to numerical identifiers and assembled into input sequences, optionally preceded by a beginning-of-sentence token, before being passed to the model to obtain next-token probability distributions. Probabilities were computed conditional upon preceding context, and surprisal was calculated as the negative logarithm of these probabilities (Levy, 2008). Word-level (i.e., ejel-level) surprisal was obtained by summing over constituent subwords, and sentence-level surprisal by aggregating across words, yielding fine-grained estimates of processing difficulty.

Human-rating data were refined by excluding responses with reading times below 1,000 ms or above 10,000 ms (data loss: 8% for datives, 6% for actives/passives, 9% for NPIs). The trimmed dataset was used for construction-specific analyses. Ordinal regression was conducted separately for each construction type using cumulative link mixed models implemented in the ordinal package (Christensen, 2023) in R (R Core Team, 2025). Fixed effects were mean-centered and deviation-coded, and all models included the maximal random-effects structure for Item and Participant supported by the design (Barr et al., 2013). A logit link function modelled the cumulative probability of higher ratings. Model fitting and validation included convergence checks, assessment of the proportional-odds assumption, and residual diagnostics. Effect sizes were estimated using Nagelkerke's R^2, derived from log-likelihood comparisons between the final model and a null model containing only an intercept for fixed effects and the full random structure. Pairwise post-hoc comparisons were performed with the emmeans package (Lenth, 2025), with *p*-values based on Wald z-tests as implemented in ordinal.

For GPT-based surprisal, we fitted separate linear mixed-effects models for each construction using lmerTest (Kuznetsova et al., 2017). The fixed-effects structure matched that of the human-rating analysis; Item was included as the sole random intercept, as multiple surprisal computations per item (proxying multiple human raters) yielded identical values. Model fit was assessed using Nakagawa's conditional R^2 (Nakagawa & Schielzeth, 2013). All other specifications mirrored those used for the human-rating analyses. Finally, we assessed alignment between human ratings and GPT surprisal via correlation analyses for each GPT variant, computing Pearson's *r*.

Results

Human rating

Table 2 summarizes human ratings by condition across the three construction types, while model outputs are reported in Table 3. For datives, acceptability differed clearly across conditions, with the Dative–Accusative pattern rated higher than the Accusative–Accusative pattern. The statistical model revealed a main effect of Condition, and Bonferroni-corrected pairwise post-hoc comparisons confirmed that Accusative–Accusative was rated significantly lower than Dative–Accusative ($p < 0.001$). These results indicate a robust preference for the Dative–Accusative construction.

[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

For actives/passives, acceptability varied substantially by condition. Human+Active, Inanimate+Active, and Human+Passive received high ratings, whereas Inanimate+Passive was rated markedly lower. The statistical model showed main effects of Animacy and Voice, as well as a significant interaction between them. Bonferroni-corrected pairwise comparisons confirmed that Inanimate+Passive was rated significantly lower than all other conditions (all *ps* < 0.001), with no reliable differences between Human+Active and Inanimate+Active, or between Human+Active and Human+Passive. Together, these findings indicate that passives are generally less acceptable than actives, and that animacy modulates acceptability specifically in the passive.

For NPIs, acceptability was high for grammatical same-clause configurations (npiE_negE; npiM_negM) and low for cross-clause configurations (npiE_negM; npiM_negE). The statistical model revealed no main effects but a significant interaction between NPI location and negation location. Bonferroni-corrected pairwise comparisons showed that same-clause conditions were rated higher than their cross-clause counterparts (npiE_negE > npiE_negM; npiM_negM > npiM_negE), with no reliable difference between the two grammatical conditions or between the two ungrammatical conditions. The results are consistent with the view that NPI licensing in Korean is not sensitive to surface adjacency per se. This is evidenced by the fact that npiM_negM received a higher rating than npiE_negM and npiM_negE.

Table 2. Descriptive statistics (human ratings)

Construction	Condition	Mean	SD
Dative	Dative–Accusative	4.77	0.62
	Accusative–Accusative	0.57	1.04
Active/ passive	Human+Active	4.58	0.92
	Human+Passive	4.23	1.20
	Inanimate+Active	4.56	0.90
	Inanimate+Passive	2.19	1.72
NPI	npiE_negE	4.22	1.38
	npiM_negM	4.43	1.07
	npiE_negM	1.53	1.50
	npiM_negE	1.61	1.75

Table 3. Statistical model outputs (human ratings)

Construction		*β*	*SE*	*z*	*p*
Dative	DAT vs ACC	–11.56	0.75	–15.5	< 0.001***
Active/ passive	Animacy	–2.15	0.48	–4.48	< 0.001***
	Voice	–3.25	0.31	–10.35	< 0.001***
	Animacy:Voice	–3.26	0.75	–4.32	< 0.001***
NPI	NPI_loc	–0.19	0.30	–0.65	0.51
	NEG_loc	0.11	0.27	0.40	0.69
	NPI_loc: NEG_loc	9.72	0.89	10.94	< 0.001***

Note. General model structure: clmm(rating ~ Condition + (1 + Condition | participant) + (1 + Condition | item). Nagelkerke's pseudo-R^2: dative = 0.37; active/passive = 0.37; NPI = 0.29. ***p < 0.001.

GPT surprisal

Table 4 summarizes GPT surprisal by model and condition across the three construction types, with full model outputs reported in Table 5. For datives, all GPT models consistently assigned lower surprisal to the Dative–Accusative condition than to the Accusative–Accusative condition. The statistical models confirmed significantly higher surprisal for Accusative–Accusative relative to Dative–Accusative, aligning with the human preference for the Dative–Accusative pattern.

For active/passive constructions, all GPT models exhibited significant main effects of Animacy and Voice. Surprisal was higher for passives than for actives and lower for inanimate than for human themes. In KoGPT-2, significant main effects of Animacy and Voice were accompanied by an interaction. Bonferroni-corrected post-hoc tests showed that Inanimate+Active had the lowest surprisal across conditions, while Human+Passive and Inanimate+Passive were significantly higher than Inanimate+Active (all *ps* < 0.001). Human+Active did not differ reliably from the passive conditions. In Ko-GPT-Trinity, main effects of Animacy and Voice were observed without an interaction. Post-hoc comparisons revealed higher surprisal for Human+Passive than Human+Active (p = 0.001), for Inanimate+Passive than Inanimate+Active (p < 0.001), and for Human+Passive than Inanimate+Active (p < 0.001).

For the NPI construction, all three models assigned higher surprisal to npiM_negM (grammatical) than to npiM_negE (ungrammatical), contrary to human ratings. In KoGPT-2, significant main effects of NPI location and negation location, as well as an interaction, were observed. Bonferroni-corrected comparisons showed npiM_negM (grammatical) > npiM_negE (ungrammatical) (p = 0.001), and npiE_negM (ungrammatical) > npiE_negE and npiM_negM (both grammatical) (all *p*s < 0.001). In Ko-GPT-Trinity, a main effect of negation location and an NPI location × negation location interaction were found. Post-hoc tests indicated that npiM_negM (grammatical) had higher surprisal than npiE_negE (grammatical) (p = 0.009) but lower surprisal than npiE_negM (ungrammatical) (all *p*s < 0.001). In mGPT, a main effect of negation location was observed. Bonferroni-corrected comparisons showed npiE_negE (grammatical) > npiE_negM (ungrammatical) (p < 0.001) and npiE_negE (grammatical) > npiM_negM (p = 0.001). Taken together, these surprisal patterns diverge from human ratings, indicating only partial and inconsistent encoding of the clausemate licensing constraint and suggesting that surprisal does not straightforwardly track perceived acceptability in this domain.

Human vs. GPT comparisons

Figure 1 illustrates correlations between human ratings and GPT surprisal ratings by condition across the three constructions. For datives, small-to-moderate correlations emerged only in Dative–Accusative (KoGPT2: r = 0.122, p = 0.015; Ko-GPT-Trinity: r = 0.065, p = 0.081), except for

mGPT; in Accusative–Accusative, correlations were weak and inconsistent across models. For actives/passives, GPT surprisal was not meaningful (all *p*s > 0.18). For the NPI construction, KoGPT-2 surprisal showed no correlation with human ratings; Ko-GPT-Trinity surprisal showed moderate negative correlations in npiE_negE ($r = 0.252$, $p = 0.048$) and npiM_negE ($r = 0.252$, $p = 0.047$).

Discussion and Conclusion

We evaluated alignment between GPT- and human-derived ratings to assess the extent to which GPT language models approximate human sentence comprehension. The results showed clear model- and construction-specific variation. Because the target constructions probe the morphosyntax–semantics interface at different levels, the models aligned more closely with morphosyntactically conditioned features (as in datives) than with features conditioned by both structure and semantics (as in actives/passives and NPIs).

For dative constructions, GPT language models captured the strong human preference for the Dative–Accusative pattern. All models assigned substantially lower surprisal to Dative–Accusative sentences than to Accusative–Accusative ones, mirroring the higher acceptability of the former. This convergence suggests that the models readily learn overt case-marking of the recipient role in the dative alternation.

For active/passive constructions, alignment was partial. Overall, surprisal was higher for passives than for actives and varied with theme animacy, indicating that the models encoded the voice contrast and treated passive voice as a more marked configuration. However, these surprisal patterns diverged from human ratings in detail. Human raters penalized passives relative to actives, but this degradation was selective and generally modest for inanimate themes, rather than uniform across conditions. By contrast, model surprisal associated animacy with increased processing cost across a wider range of conditions than those reflected in human judgements. Thus, while the models captured a broad active–passive distinction, they differed from humans in how animacy interacts with voice.

For NPI constructions, the divergence was more pronounced. Human ratings sharply distinguished grammatical from ungrammatical configurations, whereas GPT surprisal did not; in some cases, grammatical items received higher surprisal. This suggests that clause-bounded NPI licensing—a deeper structural and semantic dependency—was not consistently represented by the models, indicating that surprisal does not reliably track human judgements in this domain. One potential explanation is that the models rely disproportionately on surface-level heuristics, such as linear distance between the NPI and negation, rather than putting enough emphasis on the hierarchical constraints highlighted in human processing.

In conclusion, GPT language models appear sensitive to morphosyntactic cues such as case-marking and argument structure, while showing more limited alignment when structural features must be integrated with semantic features. These dissociations are informative for explainable AI: construction-specific comparisons can relate model outputs to established patterns in human sentence processing, clarifying which generalizations are captured by next-token prediction and where additional mechanisms may be required. Extending these benchmarks to Korean (and other lesser-studied languages) further broadens empirical coverage beyond English and provides a typologically informative setting for probing model behavior against human cognition under zero-shot evaluation.

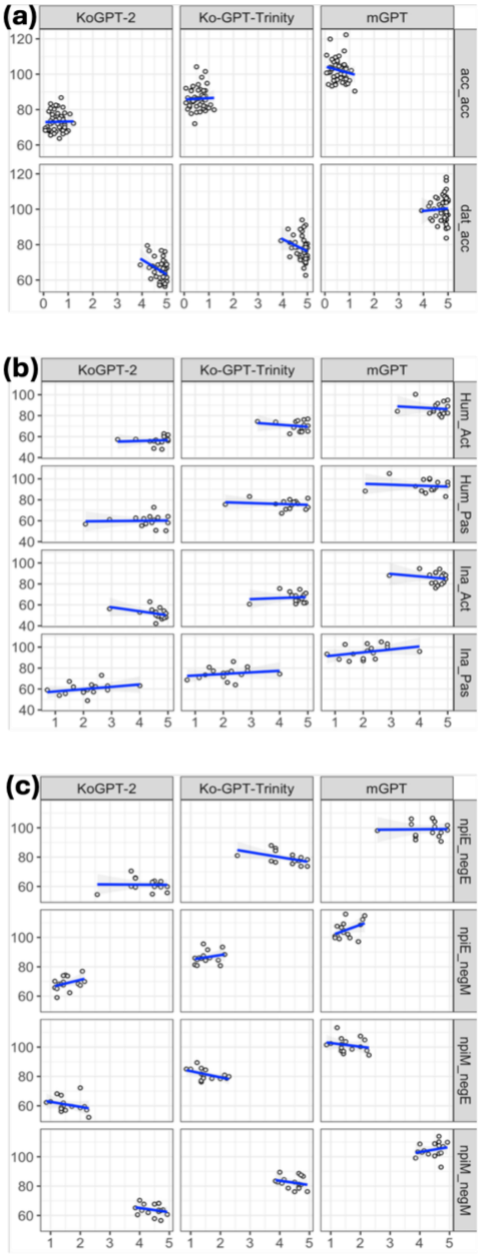

Figure 1. Correlations between human ratings and model outputs across three constructions. (a): dative; (b) active/passive; (c) NPI. X-axis: human ratings; Y-axis: surprisal. Within each subfigure, panels correspond to each model, with individual sentences as data points. Solid blue lines indicate linear regression fits, with shaded grey areas denoting 95% confidence intervals (standard errors).

Table 4. Descriptive statistics (GPT surprisal)

		KoGPT-2		Ko-GPT-Trinity		mGPT	
Construction	Condition	Mean	SD	Mean	SD	Mean	SD
Dative	Dative–Accusative	65.05	5.79	77.82	6.47	100.14	7.14
	Accusative–Accusative	73.10	5.35	86.13	6.16	102.25	6.45
Active/passive	Human+Active	56.37	4.02	70.21	4.57	86.86	5.90
	Human+Passive	60.07	5.52	75.82	4.32	93.32	5.87
	Inanimate+Active	51.87	4.97	67.08	4.81	85.93	5.70
	Inanimate+Passive	60.17	5.61	74.61	5.66	95.37	6.30
NPI	npiE_negE	61.22	4.24	79.19	4.10	98.97	5.21
	npiM_negM	63.83	3.91	86.48	4.21	105.18	5.79
	npiE_negM	68.77	4.60	81.18	3.68	101.34	4.88
	npiM_negE	60.72	5.02	82.41	4.16	104.72	5.16

Table 5. Statistical model outputs (GPT surprisal)

Model	Construction		*β*	*SE*	*t*	*p*
KoGPT-2	Dative	(intercept)	69.08	1.03	67.01	< 0.001***
		DAT vs ACC	8.05	0.95	8.48	< 0.001***
	Active/passive	(intercept)	57.12	0.88	65.14	< 0.001***
		Animacy	–2.20	1.06	–2.08	0.04*
		Voice	6.00	1.06	5.68	< 0.001***
		Animacy:Voice	4.60	2.11	2.18	0.03*
	NPI	(intercept)	63.63	0.95	66.67	< 0.001***
		NPI_loc	–2.72	0.67	–4.09	< 0.001***
		NEG_loc	5.33	0.67	8.01	< 0.001***
		NPI_loc: NEG_loc	–4.45	1.33	–3.34	0.002**
Ko-GPT-Trinity	Dative	(intercept)	81.98	1.42	57.79	< 0.001***
		DAT vs ACC	8.31	0.90	9.19	< 0.001***
	Active/passive	(intercept)	71.93	0.86	83.58	< 0.001***
		Animacy	–2.17	0.99	–2.19	0.03*
		Voice	6.57	0.99	6.62	< 0.001***
		Animacy:Voice	1.92	1.98	0.97	0.34
	NPI	(intercept)	82.32	0.83	99.46	< 0.001***
		NPI_loc	–1.04	0.67	–1.56	0.13
		NEG_loc	4.26	0.67	6.36	< 0.001***
		NPI_loc: NEG_loc	–6.06	1.34	–4.53	< 0.001***
mGPT	Dative	(intercept)	101.20	1.47	69.06	< 0.001***
		DAT vs ACC	2.11	1.02	2.06	0.04*
	Active/passive	(intercept)	90.37	1.28	70.61	< 0.001***
		Animacy	0.56	0.87	0.64	0.52
		Voice	7.95	0.87	9.09	< 0.001***
		Animacy:Voice	2.98	1.75	1.71	0.09
	NPI	(intercept)	102.55	0.97	105.83	< 0.001***
		NPI_loc	0.96	1.03	0.93	0.36
		NEG_loc	4.79	1.03	4.65	< 0.001***
		NPI_loc: NEG_loc	–2.82	2.06	–1.37	0.18

Note. General model structure: lmer(surprisal ~ Condition + (1| item). Conditional R^2: [KoGPT-2] dative = 0.55; active/passive = 0.52; NPI = 0.76. [Ko-GPT-Trinity] dative = 0.67; active/passive = 0.56; NPI = 0.70. [mGPT] dative = 0.49; active/passive = 0.76; NPI = 0.50. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

References

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., Bannard, C., Samanta, S., McCauley, S., Arnon, I., Bekman, D., Efrati, A., Berman, R., Narasimhan, B., Sharma, D. M., Nair, R. B., Fukumura, K., Campbell, S., Pye, C., Pixabaj, S. F. C., Paliz, M. M., & Mendoza, M. J. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. *Cognition, 202*, 104310. https://doi.org/10.1016/j.cognition.2020.104310

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences, 26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language, 61*(3), 374–397. https://doi.org/10.1016/j.jml.2009.07.006

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, 1–58. https://doi.org/10.1162/coli_a_00492

Cho, Y. J., & Jeon, M. G. (2015). Hankwuke swuyongseng phantanuy silhempangpeplon pikyo yenkwu [A comparative study of acceptability judgment collection methods in Korean]. *The Journal of Linguistics Science, 72*, 397–416.

Choe, H. S. (1988). *Restructuring parameters and complex predicates: A transformational approach* (Doctoral dissertation). MIT.

Choo, M., & Kwak, H.-Y. (2008). *Using Korean: A guide to contemporary usage*. Cambridge University Press.

Christensen, R. H. B. (2023). *Ordinal-regression models for ordinal data* (R package version 2023.12-4). https://CRAN.R-project.org/package=ordinal

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science, 47*(3), e13256. https://doi.org/10.1111/cogs.13256

Cuneo, N., Graves, N., Rakshit, S., & Goldberg, A. (2025). For GPT-4 as with humans: Information structure predicts acceptability of long-distance dependencies. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47). https://escholarship.org/uc/item/6rk7s243

de Varda, A. G., Marelli, M., & Amenta, S. (2024). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods, 56*(5), 5190–5213. https://doi.org/10.3758/s13428-023-02261-8

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences, 120*(51), e2309583120. https://doi.org/10.1073/pnas.2309583120

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K. A., Devinsky, O., & Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*, 369–380. https://doi.org/10.1038/s41593-022-01026-4

Haspelmath, M. (1990). The grammaticization of passive morphology. *Studies in Language, 14*(1), 25–72.

Haspelmath, M. (2015). Ditransitive constructions. *Annual Review of Linguistics, 1*(1), 19–41. https://doi.org/10.1146/annurev-linguist-030514-125204

Hawkins, R. D., Yamakoshi, T., Griffiths, T. L., & Goldberg, A. E. (2020). Investigating representations of verb bias in neural language models. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 4653–4663). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2010.02375

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans. *bioRxiv*. https://doi.org/10.1101/2022.10.04.510681

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2005.03692

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *arXiv*. https://arxiv.org/abs/2305.13264

Huang, K. J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale

benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

Jones, C. R., & Bergen, B. (2024). Does word knowledge account for the effect of world knowledge on pronoun interpretation?. *Language and Cognition*, 1–32. https://doi.org/10.1017/langcog.2024.2

Kim, H., & Shin, G-H. (2022). Effects of verb and construction frequency in sentence comprehension: A case of dative construction in Korean. *Functions of Language, 29*(3), 274–299. https://doi.org/10.1075/fol.22028.kim

Kim, H., Shin, G-H., & Hwang, H. (2020). Integration of verbal and constructional information in the second language processing of English dative constructions. *Studies in Second Language Acquisition, 42*(4), 825–847. https://doi.org/10.1017/S0272263119000743

Kuno, S. (1998). *Negative polarity items in Korean and English*. Cornell East Asia Series.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lee, S. (2024). Language model performance on English control constructions and its implications. *Journal of Linguistic Science*, *109*, 245–278. http://dx.doi.org/10.21296/jls.2024.06.109.245

Lee, S., & Schuster, S. (2022) Can language models capture syntactic associations without surface cues? A case study of reflexive anaphor licensing in English control constructions. *Society for Computation in Linguistics 5*(1), 206–211. https://doi.org/10.7275/s1kt-qg26

Lee, S., & Vu, M. (2024). The effects of distance on NPI illusive effects in BERT. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 9443–9457).

Lee, S., & Wang, S. (2023) Do language models know how to be polite? *Society for Computation in Linguistics 6*(1), 375–378. https://doi.org/10.7275/8621-5w02

Lee, Y-h. (2014). Semantic relations and multiple case constructions: an experimental approach. *Linguistic Research, 31*(2), 213–247.

Lenth, R. (2025). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.11.1).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Marvin, R., & Linzen, T. (2019). Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics, 2*(1), 373–374. https://doi.org/10.48550/arXiv.1808.09031

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.

Oh, B. D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence, 5*, 777963. https://doi.org/10.3389/frai.2022.777963

Oh, B. D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics, 11*, 336–350.

Park, S-H., & Yi, E. (2021). Perception-production asymmetry for Korean double accusative ditransitives. *Linguistic Research, 38*(1), 27–52. https://doi.org/10.17250/khisli.38.1.202103.002

Park, T. (2021). Study on the frequency and causes of the passive in English and Korean in the Gospel of John. *The Journal of Linguistics Science, 98*, 195–213.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Shin, G-H. (2020). People also avoid repetition in sentence comprehension: Evidence from multiple postposition constructions in Korean. *Linguistics Vanguard, 6*(1), 20190043. https://doi.org/10.1515/lingvan-2019-0043

Shin, G-H., & Mun, S. (2025). Modelling child comprehension: A case of suffixal passive construction in Korean. *Computer Speech & Language*, *90*, 101701. https://doi.org/10.1016/j.csl.2024.101701

Shin, G-H., Lee, S-H., & Lee, C. (2026). *Open-access dataset on acceptability ratings of Korean clausal constructions by humans and GPT models*. Proceedings of the 15th Language Resources and Evaluation Conference (LREC).

Shin, J-A., & Christianson, K. (2009). Syntactic processing in Korean–English bilingual production: Evidence from cross-linguistic structural priming. *Cognition, 112*(1), 175–180. https://doi.org/10.1016/j.cognition.2009.03.011

Shin, S-i. (2016). A study on the functions of eul/reul through examining double accusative constructions:

focusing on transitivity. *URIMALGEUL: The Korean Language and Literature, 68*, 1–35.

Shin, U., Yi, E., & Song, S. (2023). Investigating a neural language model's replicability of psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology, 14*, 937656.

Siewierska, A. (2013). Passive constructions. In M. S. Dryer & M. Haspelmath (Eds.), *WALS Online* (v2020.3). Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/107

Sohn, H. M. (1999). *The Korean language*. Cambridge University Press.

Sun, K., & Wang, R. (2025). Computational Sentence-Level Metrics of Reading Speed and Its Ramifications for Sentence Comprehension. *Cognitive Science*, *49*(7), e70092. https://doi.org/10.1111/cogs.70092

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 31st advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates, Inc.

Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1737–1743) Cognitive Science Society. https://doi.org/10.48550/arXiv.2007.06761

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretič, P., & Bowman, S. R. (2019a). Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 2877–2887).

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics, 7*, 625–641. https://doi.org/10.1162/tacl_a_00290

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 211–221). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1809.00042

Wilcox, E., Qian, P., Futrell, R., Ballesteros, M., & Levy, R. (2019). Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1).

Woo, I. H. (1997). *Wulimal phitong yenkwu [Study on a passive voice in Korean]*. Hankwukmunhwasa.

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023). The Linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1052

Yeon, J. (2003). *Korean grammatical constructions: Their form and meaning*. Saffron Books.

Yoo, M. H., Kim, J., & Song, S. (2025). Multilingual capabilities of GPT: A study of structural ambiguity. *PLoS One*, *20*(7), e0326943. https://doi.org/10.1371/journal.pone.0326943

Yoon, J. H-S. (2015). Double nominative and double accusative constructions. In L. Brown & J. Yeon (Eds.), *The Handbook of Korean Linguistics* (pp. 79–97). Oxford: John Wiley & Sons.

Zhou, X., Chen, D., Cahyawijaya, S., Duan, X., & Cai, Z. (2025, January). Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6866–6888). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.459/