Open-access Dataset on Acceptability Ratings of Korean Clausal Constructions by Humans and GPT Models

Gyu-Ho Shin[1], Soo-Hwan Lee[2], Chanyoung Lee[3]
[1]University of Illinois Chicago, ghshin@uic.edu
[2]Gyeongsang National University, soohwan.lee@gnu.ac.kr
[3]Konkuk University, clee@konkuk.ac.kr

Abstract

The present study introduces a new, open-access dataset on acceptability ratings of Korean clausal constructions at the morphosyntax–semantics interface (dative, passive, and negative polarity item). The dataset comprises (i) linguistically controlled sentence materials, (ii) ratings from targeted adult populations (individuals in their 20s), and (iii) parallel ratings from GPT variants (including ChatGPT). Alongside the release, we assess the alignment between GPT- and human-derived ratings to probe the extent to which GPT architectures can approximate patterns of human sentence comprehension. The entire dataset and code can be accessed on the OSF repository: https://doi.org/10.17605/OSF.IO/RQZU3

Keywords: GPT, human rating, clausal constructions, Korean

1. Introduction

A growing trend in language sciences is the use of computational methods to investigate linguistic phenomena. This line of research examines the ability of computational models to simulate human language behaviour (Chang, 2009; Jones and Bergen, 2024; Marvin and Linzen, 2019; Srivastava et al., 2023; Warstadt et al., 2019b; Wilcox et al., 2018), while highlighting performance differences across algorithms and/or architectural variants (Hu et al., 2020; Contreras Kallens et al., 2023; Lee and Schuster, 2022; Lee, 2024; Parrish et al., 2021a,b; Shin and Mun, 2025; Warstadt and Bowman, 2020). Recent work explores how model-based findings correspond to behavioural and corpus-based evidence that has illuminated core architectures of human language (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 and Bergen, 2024), a bias intensified by the predominance of English-based Large Language Models (LLMs). This limits the generalisability of prior findings to underrepresented languages, constraining progress towards explainable AI—that is, evaluations that sensibly and transparently relate model performance to human behaviour. The present study addresses this gap by evaluating a linking hypothesis: the extent to which LLM surprisal values and prompted ratings can serve as functional proxies for human linguistic judgements in Korean. By identifying where this link holds and where it diverges, we shed light on what constitutes an explainable AI, while acknowledging that discrepancies may be influenced by model scale and the volume of training data.

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 use 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 make robust probabilistic predictions. GPT, a decoder-only Transformer, embeds each token with positional information and processes it through stacked attention and feed-forward layers with residual connections and normalisation, 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, exhibiting zero- and few-shot generalisation through prompting (Brown et al., 2020; Radford et al., 2019).

This architecture has become central in simulating human behaviour: model probabilities and surprisal correlate with acceptability rating and processing data across tasks (Schrimpf et al., 2021; Warstadt and Bowman, 2020; Wilcox et al., 2018). Recent studies further show GPT's grammatical sensitivity and metalinguistic awareness, with GPT-4 capturing effects of information structure on acceptability (Goldstein et al., 2022; Hosseini et al., 2022; Yoo et al., 2025; Cuneo et al., 2025).

Evaluation outcomes vary across GPT families. GPT-2 closely tracks human preferences in clausal alternations and verb biases (Hawkins et al., 2020), while GPT-3 variants and ChatGPT yield mixed, prompt-sensitive results that limit reliability for grammaticality assessment (Dentella et al., 2023; Hu and Levy, 2023; Lee and Wang, 2023).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 344–356
11–16 May 2026.

In reading-time predictions, surprisal generally predicts difficulty with logarithmic cost–predictability relations; smaller GPT-2 models often outperform larger ones, which sometimes over-fit (de Varda et al., 2024; Shain et al., 2024). However, GPT-2 surprisals underestimate garden-path effects and embedded clause slowdowns, diverging from human patterns (Oh and Schuler, 2023). Large-scale benchmarks confirm such gaps: surprisal often underestimates syntactic ambiguity effects and explains only limited item-level variance, indicating that next-word prediction alone does not fully capture processing (Huang et al., 2024). Broader metrics, such as sentence surprisal or discourse-level relevance, may better capture comprehension and reading speed across languages (Sun and Wang, 2025). At the representational level, probing studies reveal that LLM activations encode subtle theoretical distinctions but with limited cross-lingual consistency, providing diagnostics of internal mechanisms (Zhou et al., 2025). Overall, GPT models demonstrate partial but qualified alignment with human sentence-comprehension patterns, with fidelity strongly shaped by size and training design.

Korean is an agglutinative, Subject–Object–Verb language characterised by overt case-marking and flexible pre-verbal argument order (Sohn, 1999). As a typologically distant language from English and underrepresented in existing research, Korean offers a valuable test case for assessing the generalisability of prior findings in the field. Currently, there is no open-access dataset of human sentence ratings in Korean, comparable to resources such as *SyntaxGym* for English (Gauthier et al., 2020).

To address this gap, we construct an open-access database of human ratings for target clausal constructions that probe the morphosyntax–semantics interface: dative construction, passive construction, and negative polarity item (NPI) construction. We then demonstrate its usage, by comparing these ratings with surprisal measures (cf. Levy 2008) computed by GPT variants, treating surprisal per sentence as a proxy for its acceptability. By releasing a well-crafted dataset and evaluating whether model predictions approximate human processing in a typologically distinct language, we expect to advance a more interpretable and nuanced account of model–human correspondence.

1.1. Target clausal constructions in Korean

1.1.1. Dative construction

Korean dative constructions appear in two main variants with the recipient-before-theme order: a Dative–Accusative pattern (1a), with a dative-marked recipient (*Jiho-eykey*[1] followed by an accusative-marked theme (*chak-ul*), and an Accusative–Accusative pattern (1b), where both recipient and theme bear accusative marking (variants *-lul/-ul* under allomorphic distribution) (Lee, 2014; Shin, 2016; Yoon, 2015).

1. Example: Dative construction in Korean

   a. Dative–Accusative
   Mina-ka Jiho-eykey chak-ul cwu-ess-ta.
   Mina-NOM Jiho-DAT book-ACC give-PST-DCL[2]
   'Mina gave a book to Jiho.'

   b. Accusative–Accusative
   Mina-ka Jiho-lul chak-ul cwu-ess-ta.
   Mina-NOM Jiho-ACC book-ACC give-PST-DCL
   'Mina gave Jiho a book.'

Although both patterns express transfer of possession (Haspelmath, 2015), corpus studies show that the Accusative–Accusative pattern is rare and dispreferred relative to the Dative–Accusative form (Cho and Jeon, 2015; Kim and Shin, 2022). While accusative case marker can signal a recipient role in restricted contexts (Shin, 2016), it more typically indicates a theme (Choo and Kwak, 2008; Sohn, 1999). Consequently, accusative-marked recipients are less favoured (Kim and Shin, 2022; Shin and Mun, 2023). Nevertheless, the Accusative–Accusative pattern is attested in communication (Park and Yi, 2021), confirming its grammaticality. Studies have used this alternation to assess acceptability judgements (Cho and Jeon, 2015; Kim and Shin, 2022; Lee, 2014) and to examine cross-linguistic priming between Korean and English (Kim et al., 2020; Shin and Christianson, 2009), supporting its validity as an alternation pair.

1.1.2. Passive construction

Passive voice, marked cross-linguistically (Haspelmath, 1990; Siewierska, 2013), is infrequent in Korean compared to its active counterpart (Park, 2021; Woo, 1997). The suffixal passive involves a nominative-marked theme subject, a dative-marked agent in an oblique position, and verbal morphology indicating passivisation (Sohn, 1999).

Active and passive constructions in Korean typically describe the same event but differ in perspec-

[1] Alternative markers include *-hanthey* [informal], *-tele/-poko* [colloquial], *-ey* [inanimate recipient], and *-kkey* [deferential intention]; Sohn, 1999)

[2] Abbreviations: ACC = accusative case marker; COMP = complementiser; DAT = dative marker; DCL = declarative ending; NEG = negator; NML = nominaliser; NOM = nominative case marker; PST = past tense marker; PSV = passive voice marker; TOP = topic marker.

tival prominence—i.e., which participant is foregrounded as grammatical subject. In actives, the agent is syntactically and semantically prominent; in passives, the theme is promoted to subject while the agent is demoted to an oblique. Prior work shows that acceptability is modulated by animacy (Sohn, 1999; Yeon, 2003): actives freely allow both [+human] and [–human] themes, whereas passives are degraded with [–human] themes. This asymmetry indicates that animacy interacts with argument structure and morphological marking in shaping the perceived naturalness of Korean passives.

2. Example: Active vs. passive construction in Korean

   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.
   Kyeunghuy-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.'

1.1.3. Negative polarity item construction

Negation can license an NPI. A sentence containing an NPI is acceptable only when the NPI and its licensor fall within the same licensing domain. This dependency is observed in English and has gained attention in the natural language processing (NLP) literature (Marvin and Linzen, 2018; Wilcox et al., 2019; Warstadt et al., 2019a; Shin et al., 2023; Lee and Vu, 2024). In Korean, sentential negation *anh* 'not' typically licenses the NPI *amwuto* 'anyone' when both occur in the same clause, which is referred to as the clausemate condition (Choe, 1988; Kuno, 1998). The clausemate condition holds in (3a) but not in (3b): in (3b) the NPI appears in the embedded clause while its licensor is in the matrix clause, rendering the sentence ungrammatical (* = ungrammatical).

3. Example: NPI construction in Korean

   a. NPI-licensing (✓ clausemate condition)
   Mina-ka amwuto o-ci ahn-ass-tako sayngkakhay-ss-ta.
   Mina-NOM anyone come-NML NEG-PST-COMP think-PST-DCL
   'Mina thought that no one came.'

   b. NPI-licensing (✗ clausemate condition)
   *Mina-ka amwuto wa-ss-tako sayngkakha-ci anh-ass-ta.
   Mina-NOM anyone come-PST-COMP think-NML NEG-PST-DCL
   'Mina did not think that anyone came.'

Making use of this dependency, we systematically vary the position of the licensor and the licensee to test how their structural configuration affects NPI-licensing and to determine the conditions under which licensing succeeds or fails across constructions.

2. Method

2.1. Dataset creation

To construct the dataset, we created 224 sentences across three constructions:

- 96 sentences for the dative construction (48 items per condition: dative–accusative, accusative–accusative)
  - *dat_acc* (Dative–Accusative): 48 sentences
  - *acc_acc* (Accusative–Accusative): 48 sentences
- 64 sentences for the active/passive construction (16 items per condition: animacy of theme subject [human, inanimate] × voice [active, passive])
  - *Hum_Act* (human theme, active voice): 16 sentences
  - *Hum_Pas* (human theme, passive voice): 16 sentences
  - *Ina_Act* (inanimate theme, active voice): 16 sentences
  - *Ina_Pas* (inanimate theme, passive voice): 16 sentences
- 64 sentences for the NPI construction (16 items per condition: NPI location [matrix clause; embedded clause] × negation location [matrix clause; embedded clause])
  - *npiM_negM* (NPI in a matrix clause, negation in a matrix clause): 16 sentences
  - *npiE_negE* (NPI in an embedded clause, negation in an embedded clause): 16 sentences
  - *npiM_negE* (NPI in a matrix clause, negation in an embedded clause): 16 sentences

– *npiE_negM* (NPI in an embedded clause, negation in a matrix clause): 16 sentences

All sentences used personal names and frequent, simple nouns for nominal arguments. Prior to the main experiment, 10 native Korean speakers normed the materials by rating their (un)grammaticality in line with the intended design. The sentences were divided into four sub-lists, randomly assigned to participants, with the order of items within each sub-list fully randomised.

For the main data collection, we recruited 56 native Korean speakers (age: M = 23.5, SD = 2.56) who completed an acceptability rating task on Qualtrics. Participants rated each sentence on a 6-point Likert scale (0 = "very unacceptable," 5 = "very acceptable"), responding as quickly as possible while maintaining accuracy. Once a response was submitted, it could not be revised. Reaction times, measured from sentence onset to scale selection, were collected to identify inattentive responses. To ensure a stable testing environment, mobile devices were not permitted. Participants completed the task in a location of their choice, provided they had a reliable internet connection and minimal distractions. The experiment took ≈ 30 minutes.

2.2. GPT surprisals

We employed two variants of GPT architecture available in Korean: KoGPT-2[3] (Korean GPT-2 model) and Ko-GPT-Trinity[4] (Korean GPT-3 model). Table 1 illustrates detailed specifications of each model. To ensure a zero-shot evaluation of grammatical sensitivity, sentences were processed in a 'closed-loop' environment without task-specific fine-tuning or few-shot prompting. All modelling was executed on a MacBook Pro (Apple M4 Max) using the Metal Performance Shaders (MPS) backend for hardware acceleration.

	KoGPT-2	Ko-GPT-Trinity
Parameter #	125M	1.2B
Layer #	12	24
Head #	12	16
Hidden layer dimension	768	2,048
FFN dimension	3,072	8,192

Table 1: Information about two GPT models.

Our workflow involved loading models from Hugging Face, standardising tokeniser behaviour, and processing sentences provided as non-empty UTF-8 strings in plain-text format (one sentence per line). Specifically, we utilised `PreTrainedTokenizerFast` for KoGPT-2 and `AutoTokenizer` for Ko-GPT-Trinity, with padding-side configurations standardised to the left to ensure consistency with decoder-only architectures. To maintain the integrity of the linguistic materials, each sentence was segmented into *ejels*[5] on whitespace, after which each *ejel* was tokenised into subword units using the model-specific tokeniser. Subword tokens were mapped to numerical identifiers and assembled into input sequences, optionally with a beginning-of-sentence token, before being passed to the model to obtain next-token probability distributions.

For each model, probabilities were assigned to tokens conditional on their preceding context. Surprisal was then computed as the negative logarithm of these conditional probabilities (Levy, 2008). Word-level surprisal (i.e., *ejel*-level) was derived by summing the surprisals of constituent subword tokens, while sentence-level surprisal was obtained by aggregating these values across all words in the sequence. This approach allows us to evaluate a linking hypothesis: the extent to which these raw computational metrics—free from the stochastic variability of natural language prompts—can serve as a transparent proxy for human acceptability judgements, providing fine-grained profiles of processing difficulty across sentences.

2.3. ChatGPT rating

We also evaluated ChatGPT-5's sentence-level rating performance (accessed on 25 September 2025) using controlled prompting that closely mirrored the instructions given to human raters. For each item, we collected a numeric score and a one-sentence rationale justifying the judgement. Two human annotators administered the prompts independently, with ChatGPT's memory left unchanged and not updated across trials.

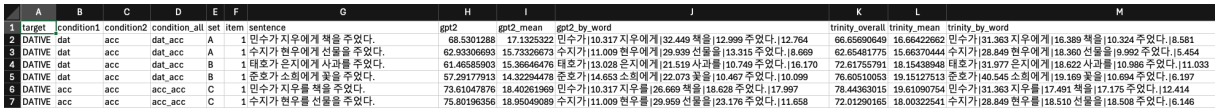
The Korean prompt used in the experiment is as follows: "아래에 주어지는 문장이 한국어에서 얼마나 자연스러운지를 0부터 5까지 6점 척도로 판단해 줘. 0은 '매우 부자연스러움', 5는 '매우 자연스러움'이야. 정수(0, 1, 2, 3, 4, 5)로 판단을 하면서, 왜 그렇게 판단을 했는지 이유를 한 문장으로 말해줘." The English translation of the prompt is as follows: 'Rate how natural the following sentence is 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.'

[3] https://huggingface.co/skt/kogpt2-base-v2

[4] https://huggingface.co/skt/ko-gpt-trinity-1.2B-v0.5

[5] An *ejel*, comprising one or more morphemes, is a Korean writing unit typically separated by spaces. For example, *na-nun* 'I-TOP' in *na-nun khi-ka khu-ta* 'I-TOP height-NOM tall-DCL' ('I am tall.') is one *ejel* .

(a) rating_combined



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	target	condition1	condition2	condition_all	set	item	sentence	gpt2	gpt2_mean	gpt2_by_word		trinity_overall	trinity_mean	trinity_by_word		
2	DATIVE	dat	acc	dat_acc	A	1	민수가 지우에게 책을 주었다.	68.5301288	17.1325322	민수가|10.317 지우에게|32.449 책을|12.999 주었다.|12.764		66.65690649	16.66422662	민수가|31.363 지우에게|16.389 책을|10.324 주었다.|8.581		
3	DATIVE	dat	acc	dat_acc	A	1	수지가 현우에게 선물을 주었다.	62.93306693	15.73326673	수지가|11.009 현우에게|29.939 선물을|13.315 주었다.|8.669		62.65481775	15.66370444	수지가|28.849 현우에게|18.360 선물을|9.992 주었다.|5.454		
4	DATIVE	dat	acc	dat_acc	B	1	태호가 은지에게 사과를 주었다.	61.46585903	15.36646476	태호가|13.028 은지에게|21.519 사과를|10.749 주었다.|16.170		72.61755791	18.15438948	태호가|31.977 은지에게|18.622 사과를|10.986 주었다.|11.033		
5	DATIVE	dat	acc	dat_acc	B	1	준호가 소희에게 꽃을 주었다.	57.29177913	14.32294478	준호가|14.653 소희에게|22.073 꽃을|10.467 주었다.|10.099		76.60510053	19.15127513	준호가|40.545 소희에게|19.169 꽃을|10.694 주었다.|6.197		
6	DATIVE	acc	acc	acc_acc	C	1	민수가 지우를 책을 주었다.	73.61047876	18.40261969	민수가|10.317 지우를|26.669 책을|18.628 주었다.|17.997		78.44363015	19.61090754	민수가|31.363 지우를|17.491 책을|17.175 주었다.|12.414		
7	DATIVE	acc	acc	acc_acc	C	1	수지가 현우를 선물을 주었다.	75.80196356	18.95049089	수지가|11.009 현우를|29.959 선물을|23.176 주었다.|11.658		72.01290165	18.00322541	수지가|28.849 현우를|18.510 선물을|18.508 주었다.|6.146		

(b) rating_gpt

	A	B	C	D	E	F	G	H	I	J
1	target	participant	condition1	condition2	condition_all	set	item	sentence	rating	rt
2	NPI	1	npiE	negE	npiE_negE	A	1	민서가 아무도 오지 않았다고 판단했다.	4	5.041
3	NPI	1	npiM	negM	npiM_negM	A	2	아무도 준우가 갔다고 생각하지 않았다.	5	4.532
4	NPI	1	npiE	negM	npiE_negM	A	3	민재가 아무도 죽었다고 확신하지 않았다.	1	3.838
5	NPI	1	npiM	negE	npiM_negE	A	4	아무도 영희가 멈추지 않았다고 지적했다.	0	8.986
6	NPI	1	npiE	negE	npiE_negE	A	5	준서가 아무도 일어나지 않았다고 언급했다.	5	5.226
7	NPI	1	npiM	negM	npiM_negM	A	6	아무도 윤서가 뛰었다고 주장하지 않았다.	5	4.793

(c) rating_human

Figure 1: Structure of the final dataset.

2.4. Final data structure

Figure 1 illustrates the overall structure of the dataset. The resource is distributed as an openly accessible .xlsx workbook, designed to be compatible with standard spreadsheet applications to ensure maximum usability and transparency. This format allows seamless access not only for researchers specialising in NLP but also for scholars from related disciplines, educators, and practitioners who may lack technical expertise in computational methods.

The workbook consists of three well-structured tabs, each catering to distinct aspects of the data. The first tab, `rating_combined` (Figure 1(a)), presents by-sentence information including human ratings (mean, standard deviation, standard error), GPT surprisals (per-model means), and ChatGPT ratings (numerical scores accompanied by concise one-sentence rationales that two prompters obtained from the model). It also incorporates coded metadata for construction type (DATIVE, ACTPSV, NPI), experimental condition, and item identifiers, facilitating reproducibility and secondary analysis. The second tab, `rating_gpt` (Figure 1(b)), contains GPT surprisal values and ChatGPT ratings exclusively, providing by-word surprisal measurements and per-sentence means to enable fine-grained computational comparison. The third tab, `rating_human` (Figure 1(c)), contains human-derived data only, including coded participant identifiers and per-sentence reaction times.

Collectively, this open-access format ensures that the dataset remains transparent, reusable, and conducive to inclusive research engagement across disciplinary and methodological boundaries.

3. Sample analysis: Human ratings vs. GPT-based ratings

Using the dataset, we conducted sample analysis that assessed alignment between human ratings and GPT surprisals to gauge how closely GPT architectures approximate patterns of human sentence comprehension.

3.1. Statistical analysis

For the human sentence ratings, raw data were refined by excluding responses with reading times below 1,000 ms or above 10,000 ms (data loss: 8.18% for the dative construction; 6.47% for the passive construction; 8.59% for the NPI construction). The trimmed dataset was then used for construction-specific analyses.

Ordinal regression per construction type was conducted with cumulative link mixed models using the *ordinal* package (Christensen, 2023) in R (R Core Team, 2025). Fixed effects (i.e., conditions) were mean-centred and deviation-coded, and all models included the maximal random-effects structure with *Item* and *Participant*, as 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 via 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 using the *emmeans* package (Lenth, 2025), with *p*-values based on Wald *z*-tests as implemented in *ordinal*.

Dative construction				Active/passive construction				NPI construction			
Condition	Mean	SD	SE	Condition	Mean	SD	SE	Condition	Mean	SD	SE
dat_acc	4.774	0.623	0.025	Hum_Act	4.585	0.917	0.063	npiE_negE	4.223	1.376	0.097
acc_acc	0.574	1.041	0.042	Ina_Act	4.556	0.898	0.063	npiM_negM	4.426	1.069	0.075
				Hum_Pas	4.229	1.202	0.082	npiE_negM	1.532	1.500	0.105
				Ina_Pas	2.188	1.720	0.119	npiM_negE	1.608	1.751	0.121

Table 2: Descriptive statistics: Human ratings by condition across three constructions.

Dative construction					Active/passive construction					NPI construction				
	Estimate	SE	*z*	*p*		Estimate	SE	*z*	*p*		Estimate	SE	*z*	*p*
DATvsACC	-11.559	0.746	-15.5	0.001***	Animacy	-2.154	0.481	-4.480	0.001***	NPI_loc	-0.193	0.296	-0.654	0.513
					Voice	-3.253	0.314	-10.347	0.001***	NEG_loc	0.107	0.271	0.395	0.693
					Animacy:Voice	-3.262	0.754	-4.324	0.001***	NPI_loc:NEG_loc	9.718	0.889	10.938	0.001***

Table 3: Statistical model outputs: Human ratings by condition across three constructions. General model structure (Ordinal regression modelling with random effects): `clmm(rating ~ condition + (1 + condition | participant) + (1 + condition | item), data = ajt, link = "logit")`. Nagelkerke's pseudo R^2: Dative = 0.367; Passive = 0.367; NPI = 0.289. $^{***}p < 0.001$.

For GPT-based sentence surprisals, we fitted separate linear mixed-effects models for each construction using *lme4* (Bates et al., 2015). The fixed-effects structure matched 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. Each model's R^2 was computed using Nakagawa's R^2 (Nakagawa and Schielzeth, 2013), which considers both fixed and random effects (conditional R^2). All other specifications mirrored those used for the human sentence rating data. Finally, to compare human ratings and GPT surprisals, we conducted correlation analyses for each GPT variant, also computing Pearson's *r*.

For ChatGPT sentence ratings, the target statistical model could not be fitted because the data exhibited quasi-complete separation: ratings were near-deterministically associated with condition (especially for dative and active/passive constructions), resulting in non-identifiable parameter estimates and infinite likelihood. Accordingly, rather than mirroring the complex mixed-effects modelling used for human rating data, we focused on descriptive statistics and correlations with human ratings. Alongside numeric scores, we examined the one-sentence rationales by condition to understand how ChatGPT determined acceptability.

3.2. Human rating

Table 2 presents a summary of human ratings by condition across three construction type; model outputs are summarised in Table 3. For the dative construction, acceptability differed clearly between conditions, with Dative–Accusative rated higher than Accusative–Accusative. The statistical model yielded a main effect of *Condition*; Bonferroni-corrected pairwise post-hoc comparisons further revealed that Accusative–Accusative was rated significantly lower than Dative–Accusative ($p <$ 0.001). These findings indicate a preference for the Dative–Accusative condition over the Accusative–Accusative condition.

For the passive construction, acceptability varied markedly across conditions: Human_Active, Inanimate_Active, and Human_Passive were rated highly, whereas Inanimate_Passive received substantially lower ratings. The statistical model showed main effects of *Animacy* and *Voice*, plus an interaction between the two. Bonferroni-corrected pairwise post-hoc comparisons confirmed Inanimate_Passive was rated significantly lower than all the other conditions (all *p*s < 0.001), with no meaningful differences between Human_Active and Inanimate_Active, or between Human_Active and Human_Passive. These results indicate that passives are generally rated lower than actives, and that animacy modulates acceptability only in the passive.

For the NPI construction, 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 yielded no main effects but a significant *NPI location* $\times$ *NEG location* interaction. Bonferroni-corrected pairwise post-hoc comparisons further revealed 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, nor between the two ungrammatical conditions. These results support a clausemate licensing requirement: an NPI is acceptable when the NPI and its licensor are placed in the same clause but degraded otherwise.

3.3. GPT surprisal

Table 4 presents a summary of GPT surprisals by model and condition across three construction type; model outputs are summarised in Table 5. For the dative construction, both GPT mod-

Dative construction						Active/passive construction						NPI construction					
Model	Condition	N	Mean	SD	SE	Model	Condition	N	Mean	SD	SE	Model	Condition	N	Mean	SD	SE
GPT2	dat_acc	48	65.053	5.786	0.835	GPT2	Hum_Act	16	56.365	4.020	1.005	GPT2	npiE_negE	16	61.217	4.244	1.061
	acc_acc	48	73.103	5.348	0.772		Ina_Act	16	51.866	4.966	1.242		npiM_negM	16	63.827	3.909	0.977
GPT-Trinity	dat_acc	48	77.822	6.470	0.934		Hum_Pas	16	60.069	5.522	1.381		npiE_negM	16	68.774	4.595	1.149
	acc_acc	48	86.129	6.165	0.890		Ina_Pas	16	60.167	5.611	1.403		npiM_negE	16	60.720	5.016	1.254
						GPT-Trinity	Hum_Act	16	70.213	4.565	1.141	GPT-Trinity	npiE_negE	16	79.191	4.100	1.025
							Ina_Act	16	67.084	4.810	1.203		npiM_negM	16	82.408	4.157	1.039
							Hum_Pas	16	75.824	4.316	1.079		npiE_negM	16	86.482	4.208	1.052
							Ina_Pas	16	74.613	5.663	1.416		npiM_negE	16	81.182	3.681	0.920

Table 4: Descriptive statistics: GPT model ratings by condition across three constructions.

Dative construction					Active/passive construction					NPI construction				
	Estimate	SE	*t*	*p*		Estimate	SE	*t*	*p*		Estimate	SE	*t*	*p*
GPT-2														
(Intercept)	69.078	1.031	67.008	0.001***	(Intercept)	57.117	0.877	65.139	0.001***	(Intercept)	63.635	0.955	66.665	0.001***
DATvsACC	8.049	0.949	8.478	0.001***	Animacy	-2.200	1.057	-2.082	0.043*	NPI_loc	-2.722	0.665	-4.091	0.001***
					Voice	6.002	1.057	5.680	0.001***	NEG_loc	5.332	0.665	8.014	0.001***
					Animacy:Voice	4.598	2.113	2.176	0.03*	NPI_loc:NEG_loc	-4.450	1.331	-3.344	0.002*
GPT-Trinity														
(Intercept)	81.976	1.419	57.786	0.001***	(Intercept)	71.934	0.861	83.581	0.001***	(Intercept)	82.316	0.828	99.457	0.001***
DATvsACC	8.307	0.904	9.188	0.001***	Animacy	-2.170	0.992	-2.186	0.034*	NPI_loc	-1.041	0.669	-1.556	0.127
					Voice	6.570	0.992	6.620	0.001***	NEG_loc	4.259	0.669	6.363	0.001***
					Animacy:Voice	1.918	1.985	0.966	0.339	NPI_loc:NEG_loc	-6.064	1.338	-4.530	0.001***

Table 5: Statistical model outputs for GPT-2 and GPT-Trinity ratings across three constructions. General model structure (Linear mixed-effects modelling): `lmer(surprisal ~ condition + (1 | item), data = ajt, REML = TRUE)`; the maximal random-effects structure did not converge. Conditional R^2: [GPT-2] Dative = 0.550; Passive = 0.523; NPI = 0.766. [GPT-Trinity] Dative = 0.666; Passive = 0.563; NPI = 0.696. $^{*}p < 0.05$; $^{***}p < 0.001$.

els consistently yielded lower surprisals in Dative–Accusative than in Accusative–Accusative. The statistical models confirmed significantly higher surprisals in Accusative–Accusative relative to Dative–Accusative. This aligns with the human-rating preference for Dative–Accusative.

For the passive construction, both GPT models showed significant main effects of *Animacy* and *Voice*. Surprisal was higher for passives than actives and lower for inanimate than human patients. In GPT-2, there were main effects of *Animacy* and *Voice* and a significant interaction between them. Bonferroni-corrected pairwise post-hoc comparisons showed that Inanimate_Active had the lowest surprisal amongst all conditions, whereas Human_Passive and Inanimate_Passive were significantly higher than Inanimate_Active (all *p*s < 0.001). Human_Active did not differ significantly from the passives. In GPT-Trinity, main effects of *Animacy* and *Voice* were found without an interaction. Bonferroni-corrected pairwise post-hoc comparisons revealed higher surprisals 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 ($p < 0.001$).

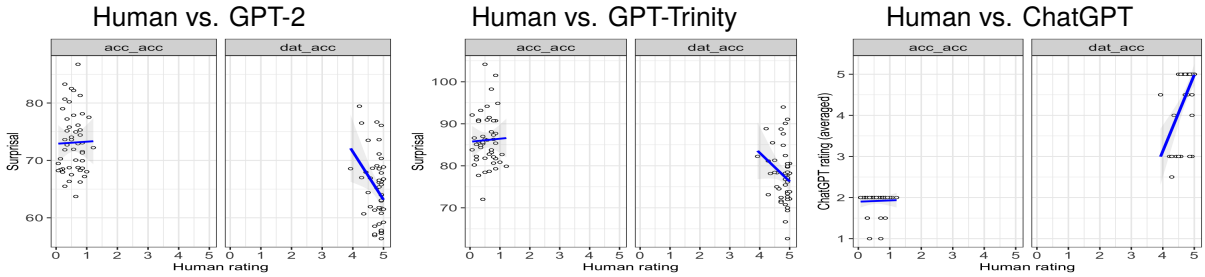
For the NPI construction, both GPT-2 and GPT-Trinity assigned higher surprisals to npiM_negM (grammatical) than to npiM_negE (ungrammatical), which runs counter to our findings from human ratings. In GPT-2, the model showed main effects of *NPI location* and *NEG location*, plus an interaction between the two; Bonferroni-corrected pairwise post-hoc comparisons indicated 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 GPT-Trinity, there was a main effect of *NEG location* and an *NPI location* × *NEG location* interaction; Bonferroni-corrected pairwise post-hoc comparisons showed npiM_negM (grammatical) > npiE_negE (grammatical) ($p = 0.009$) but < npiE_negM (ungrammatical) (*p*s < 0.001). Taken together, GPT surprisals diverged from human ratings, suggesting only a partial and inconsistent encoding of the clausemate licensing constraint. This implies that surprisal may not straightforwardly track perceived acceptability in this domain.

3.4. ChatGPT rating

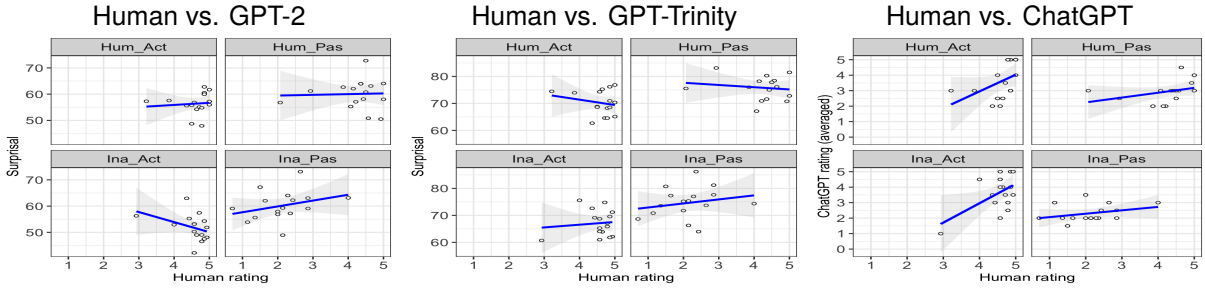
Table 6 presents ChatGPT ratings by condition across three construction types. For the dative construction, Dative–Accusative received higher ratings than Accusative–Accusative, consistent with human ratings and GPT surprisals; ChatGPT's one-sentence rationales likewise favoured Dative–Accusative. For the passive construction, ChatGPT showed a clear voice effect, with passives rated lower than actives; within passives, Human_Passive was rated slightly higher than Inan-

Dative construction					Active/passive construction					NPI construction				
Condition	N	Mean	SD	SE	Condition	N	Mean	SD	SE	Condition	N	Mean	SD	SE
dat_acc	96	4.562	0.792	0.081	Hum_Act	32	3.594	1.132	0.200	npiE_negE	32	3.531	0.621	0.110
acc_acc	96	1.917	0.345	0.035	Ina_Act	32	3.625	1.184	0.209	npiM_negM	32	4.062	0.716	0.126
					Hum_Pas	32	2.937	0.716	0.126	npiE_negM	32	2.375	0.553	0.098
					Ina_Pas	32	2.312	0.592	0.105	npiM_negE	32	2.719	0.457	0.081

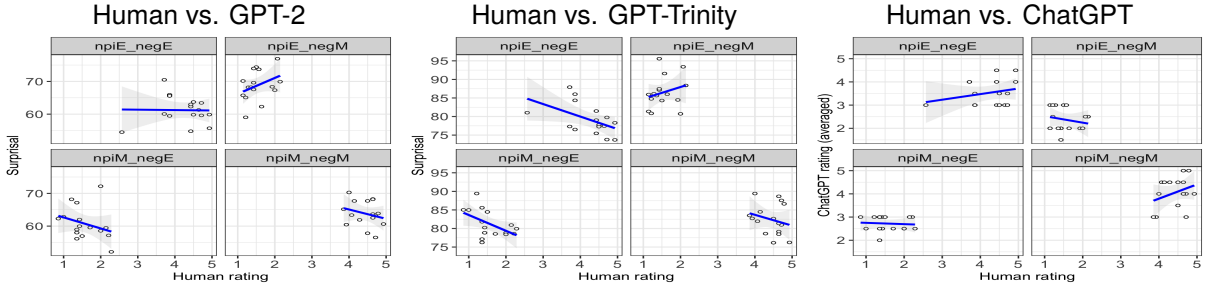
Table 6: Descriptive statistics: ChatGPT ratings by condition across three constructions.



(a) Dative construction



(b) Active/passive construction



(c) NPI construction

Figure 2: Correlation per condition between human ratings and model ratings across three constructions. Within each subfigure, the three panels show GPT-2, GPT-Trinity, and ChatGPT, respectively (sentences as data points). Within each subfigure, a solid blue line represents a regression line, and a grey area indicates a 95% confidence interval.

imate_Passive, indicating a limited animacy effect. For the NPI construction, npiE_negE and npiM_negM (grammatical) received higher ratings than npiE_negM and npiM_negE (ungrammatical); the accompanying rationales broadly paralleled the human-rating patterns.

3.5. Human vs. GPT comparisons

Figure 2 illustrates correlations between human ratings and GPT surprisals/ChatGPT ratings by condition across the three constructions. For the dative construction, small-to-moderate correlations emerged only in Dative–Accusative (GPT2: r = 0.122, p = 0.015***; GPT-Trinity: r = 0.065, p = 0.081^{ms}; ChatGPT: r = 0.335, $p < 0.001^{***}$); in Accusative–Accusative, correlations were weak and

inconsistent across models. For the passive construction, GPT surprisals were not meaningful (all *p*s > 0.1), whereas ChatGPT showed a moderate positive correlation in Inanimate_Active (r = 0.54, p = 0.03). For the NPI construction, GPT-2 surprisals showed no correlation with human ratings; GPT-Trinity surprisals showed moderate negative correlations in npiE_negE (r = 0.252, p = 0.048) and npiM_negE (r = 0.252, p = 0.047).

4. Conclusion

In the current work, we released an open-access dataset of acceptability ratings for Korean clausal constructions that manifest the morphosyntax–semantics interface (i.e., dative, passive, and NPI constructions). The dataset comprised (i) linguistically controlled sentence materials, (ii) ratings from targeted adult populations, and (iii) parallel ratings from multiple GPT variants.

Alongside the release, we evaluated alignment between GPT- and human-derived ratings to gauge how closely GPT architectures approximate human sentence comprehension. Results revealed model- and construction-specific variation. Specifically, given the salient morphosyntax–semantics interface involving our target constructions, GPT appeared to align more closely with structurally conditioned features (e.g., those found in datives) and less with semantically conditioned features (e.g., those found in passives or NPIs).

The decision to use sentence-level surprisal as the primary metric aimed to mitigate the 'prompt-sensitivity' often reported in LLMs, where minor phrasing variations yield inconsistent grammaticality assessments. By focusing on the probability distributions inherent in the models' pre-training, we sought a more direct measurement of internalised grammatical knowledge; however, results suggest this approach only partially aligns with human linguistic judgements. Despite the aforementioned caveats regarding model scale and training data coverage, our findings further imply that surprisal may not consistently support a transparently explainable AI for complex Korean morphosyntax or serve as a robust component of the linking hypothesis in this context.

While modest in size relative to large-scale resources in major languages—such as *SyntaxGym* for English (Gauthier et al., 2020)—this dataset offers a concrete, empirically grounded starting point for downstream NLP research on Korean, helping to mitigate the field's English-centric bias and to improve data accessibility for researchers.

5. Extra space for ethical considerations and limitations

Informed consent was obtained from all participants, and the study protocol was approved by the Institutional Review Board at the University of Illinois Chicago (approval number: STUDY2023-1271) to ensure compliance with ethical research standards. The current work has its limitations. The dataset was relatively small and the construction types were constrained compared to large-scale resources in major languages. Although no comparable open-access Korean dataset exists, we acknowledge these limits and plan to expand both size and coverage in future work. Regarding ChatGPT, the two prompters obtained slightly different ratings for the same sentences, suggesting sensitivity to extraneous factors (e.g., subtle prompt-phrasing or formatting differences, stochastic decoding, session/state variability). Researchers may therefore wish to use multiple prompters and repeated runs under tightly controlled prompting environments (e.g., fixed templates and decoding parameters), and to empirically examine how such controls influence ChatGPT's performance.

6. References

B. Ambridge, R. Maitreyee, T. Tatsumi, L. Doherty, S. Zicherman, P. M. Pedro, C. Bannard, S. Samanta, S. McCauley, I. Arnon, D. Bekman, A. Efrati, R. Berman, B. Narasimhan, D. M. Sharma, R. B. Nair, K. Fukumura, S. Campbell, C. Pye, S. F. C. Pixabaj, M. M. Paliz, and M. J. Mendoza. 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.

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

D. Bates, M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

D. E. Blasi, J. Henrich, E. Adamou, D. Kemmerer, and A. Majid. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh,

D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

F. Chang. 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.

T. A. Chang and B. K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, pages 1–58.

Y. J. Cho and M. G. Jeon. 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.

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

M. Choo and H-Y. Kwak. 2008. *Using Korean: A Guide to Contemporary Usage*. Cambridge University Press, Cambridge.

R. H. B. Christensen. 2023. *Ordinal-Regression Models for Ordinal Data*. R package version 2023.12-4.

P. Contreras Kallens, R. D. Kristensen-McLachlan, and M. H. Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, 47(3):e13256.

N. Cuneo, N. Graves, S. Rakshit, and A. Goldberg. 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*, volume 47.

A. G. de Varda, M. Marelli, and S. Amenta. 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.

V. Dentella, F. Günther, and E. Leivada. 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.

J. Gauthier, J. Hu, E. Wilcox, P. Qian, and R. 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. Association for Computational Linguistics.

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

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

M. Haspelmath. 2015. Ditransitive constructions. *Annual Review of Linguistics*, 1(1):19–41.

R. D. Hawkins, T. Yamakoshi, T. L. Griffiths, and A. E. Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4653–4663. Association for Computational Linguistics.

E. A. Hosseini, M. Schrimpf, Y. Zhang, S. Bowman, N. Zaslavsky, and E. Fedorenko. 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv*.

J. Hu, J. Gauthier, P. Qian, E. Wilcox, and R. P. Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the Association for Computational Linguistics*, pages 1725–1744. Association for Computational Linguistics.

J. Hu and R. Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

K. J. Huang, S. Arehalli, M. Kugemoto, C. Muxica, G. Prasad, B. Dillon, and T. Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

C. R. Jones and B. Bergen. 2024. Does word knowledge account for the effect of world knowledge on pronoun interpretation? *Language and Cognition*, pages 1–32.

H. Kim and G-H. Shin. 2022. Effects of verb and construction frequency in sentence comprehension: A case of dative construction in Korean. *Functions of Language*, 29(3):274–299.

H. Kim, G-H. Shin, and H. Hwang. 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.

S. Kuno. 1998. Negative polarity items in Korean and English. *Cornell East Asia Series*, 98:87–132.

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

S-H. Lee. 2024. Language model performance on English control constructions and its implications. *Journal of Linguistic Science*, pages 245–278.

S-H. Lee and S. 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.

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

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

R. Lenth. 2025. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.1.

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

R. Marvin and T. Linzen. 2018. Targeted syntactic evaluation of language models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

R. Marvin and T. Linzen. 2019. Targeted syntactic evaluation of language models. In *Proceedings of the Society for Computation in Linguistics*, volume 2, pages 373–374.

S. Nakagawa and H. Schielzeth. 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.

B. D. Oh, C. Clark, and W. Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

B. D. Oh and W. Schuler. 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.

S-H. Park and E. Yi. 2021. Perception-production asymmetry for Korean double accusative ditransitives. *Linguistic Research*, 38(1):27–52.

T. Park. 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.

A. Parrish, W. Huang, O. Agha, S-H. Lee, N. Nangia, A. Warstadt, K. Aggarwal, E. Allaway, T. Linzen, and S. R. Bowman. 2021a. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A. Parrish, S. Schuster, A. Warstadt, O. Agha, S-H. Lee, Z. Zhao, S. R. Bowman, and T. Linzen. 2021b. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

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

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

G-H. Shin and S. Mun. 2023. Korean-speaking children's constructional knowledge about a transitive event: Corpus analysis and Bayesian modelling. *Journal of Child Language*, 50(2):311–337.

G-H. Shin and S. Mun. 2025. Modelling child comprehension: A case of suffixal passive construction in Korean. *Computer Speech & Language*, 90:101701.

J-A. Shin and K. Christianson. 2009. Syntactic processing in Korean–English bilingual production: Evidence from cross-linguistic structural priming. *Cognition*, 112(1):175–180.

S-i. Shin. 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.

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

A. Siewierska. 2013. Passive constructions. In M. S. Dryer and M. Haspelmath, editors, *WALS Online (v2020.3)*. Max Planck Institute for Evolutionary Anthropology.

H. M. Sohn. 1999. *The Korean Language*. Cambridge University Press, Cambridge.

A. Srivastava, A. Rastogi, A. Rao, A. A. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

K. Sun and R. Wang. 2025. Computational sentence-level metrics of reading speed and its ramifications for sentence comprehension. *Cognitive Science*, 49(7):e70092.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 31, pages 5998–6008. Curran Associates, Inc.

A. Warstadt and S. R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 1737–1743. Cognitive Science Society.

A. Warstadt, Y. Cao, I. Grosu, W. Peng, H. Blix, Y. Nie, A. Alsop, S. Bordia, H. Liu, A. Parrish, S-F. Wang, J. Phang, A. Mohananey, P. M. Htut, P. Jeretič, and S. R. Bowman. 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 (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

A. Warstadt, A. Singh, and S. R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

E. Wilcox, R. Levy, T. Morita, and R. 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. Association for Computational Linguistics.

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

I. H. Woo. 1997. *wulimal phitong yenkwu [Study on a passive voice in Korean]*. Hankwukmwunhwasa, Seoul.

W. Xu, J. Chon, T. Liu, and R. Futrell. 2023. The linearity of the effect of surprisal on reading times across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721. Association for Computational Linguistics.

J. Yeon. 2003. *Korean Grammatical Constructions: Their Form and Meaning*. Saffron Books, London.

M. H. Yoo, J. Kim, and S. Song. 2025. Multilingual capabilities of GPT: A study of structural ambiguity. *PLoS One*, 20(7):e0326943.

J. H-S. Yoon. 2015. Double nominative and double accusative constructions. In L. Brown and J. Yeon, editors, *The Handbook of Korean Linguistics*, pages 79–97. John Wiley Sons, Oxford.

X. Zhou, D. Chen, S. Cahyawijaya, X. Duan, and Z. Cai. 2025. Linguistic minimal pairs elicit linguistic similarity in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888. Association for Computational Linguistics.