Received: 20 August 2022 | Revised: 10 March 2023 | Accepted: 8 April 2023

DOI: 10.1111/desc.13405

Developmental Science WILEY

RESEARCH ARTICLE

Explainability of neural networks for child language: Agent-First strategy in comprehension of Korean active transitive construction

Gyu-Ho Shin[1,2] | Seongmin Mun[3]

[1]Department of Linguistics, University of Illinois Chicago, Chicago, IL, USA

[2]Department of Asian Studies, Palacky University Olomouc, Olomouc, Czech Republic

[3]Humanities Research Institute, Ajou University, Suwon-si, Gyeonggi-do, South Korea

Correspondence
Gyu-Ho Shin, Department of Linguistics, University of Illinois Chicago, 601 S Morgan St, Chicago, IL 60607, USA and Department of Asian Studies, Palacky University Olomouc, tř. Svobody 26, 779 00 Olomouc, Czech Republic.
Email: gyuhoshin@gmail.com

Funding information
European Regional Development Fund, Grant/Award Number: CZ.02.1.01/0.0/0.0/16_019/0000791; National Research Foundation of Korea, Grant/Award Number: NRF-2022S1A5C2A02090368

Abstract

This study investigates how neural networks address the properties of children's linguistic knowledge, with a focus on the *Agent-First* strategy in comprehension of an active transitive construction in Korean. We develop various neural-network models and measure their classification performance on the test stimuli used in a behavioural experiment involving scrambling and omission of sentential components at varying degrees. Results show that, despite some compatibility of these models' performance with the children's response patterns, their performance does not fully approximate the children's utilisation of this strategy, demonstrating by-model and by-condition asymmetries. This study's findings suggest that neural networks can utilise information about formal co-occurrences to access the intended message to a certain degree, but the outcome of this process may be substantially different from how a child (as a developing processor) engages in comprehension. This implies some limits of neural networks on revealing the developmental trajectories of child language.

KEYWORDS
active transitive, *Agent-First* strategy, child comprehension, Korean, neural network

Research Highlights

- This study investigates how neural networks address properties of child language.
- We focus on the *Agent-First* strategy in comprehension of Korean active transitive.
- Results show by-model/condition asymmetries against children's response patterns.
- This implies some limits of neural networks on revealing properties of child language.

1 | INTRODUCTION

There is growing interest in the ways neural networks (NNs) address human language behaviour (Futrell & Levy, 2019; Hawkins et al., 2020; Hu et al., 2020; Warstadt & Bowman, 2020). Artificial NNs, analogous to biological NNs in human brains (Haykin, 2009; Hopfield, 1982; Jordan, 1997), are proposed as a computing system which comprises weighted and layered interconnections amongst processing units (loosely modelling neurons in the brain) responding to input in parallel and producing output through propagation (see Goldberg, 2017; Kriesel, 2007 for in-depth descriptions of NNs). The continuous development of NN algorithms in computer science has made their

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Developmental Science* published by John Wiley & Sons Ltd.

internal mechanisms deviate from how biological neurons operate in the human brain (e.g., Crick, 1989), but NNs have been applied to various disciplines in reality (Abiodun et al., 2018). Specifically in the literature on language development, while researchers adopt various computational modelling techniques to reveal developmental trajectories of linguistic knowledge (Alishahi & Stevenson, 2008; Ambridge et al., 2020; Bannard et al., 2009; Chang, 2009; Divjak et al., 2021; You et al., 2021), the current research practice bears two major caveats. First, findings are based exclusively on a limited range of languages such as English, generating a sampling bias towards those languages and populations speaking those languages (cf. Kidd & Garcia, 2022; Nielsen et al., 2017). Second, compared to an emerging strand of literature targeting adult language (Hawkins et al., 2020; Hu et al., 2020; Oh et al., 2022; Warstadt & Bowman, 2020; Warstadt et al., 2019), there is little research on how NNs approximate the characteristics of child language found in corpus analysis and/or behavioural experiments.

The current study attempts to fill these gaps, inquiring into the extent to which NNs capture the properties of children's linguistic knowledge for language other than English, with a special focus on the *Agent-First* strategy in comprehension. Children often map the first noun (mostly the subject) of a sentence to an agent role in comprehension. This strategy, whether it be a temporary bias in online processing (e.g., Abbot-Smith et al., 2017) or a heuristic persistent over the entire comprehension (e.g., Slobin & Bever, 1982), is driven from various sources. To illustrate, repeated exposure to the particular association between the first argument and agenthood provides a prototype for thematic role ordering (Bates & MacWhinney, 1989). The first item in a sequence also holds a privileged status in human cognition. Language users employ the first element in a sentence as a starting point for language behaviour, which guides the rest of the sentence (MacWhinney, 1977). When comprehenders initiate linguistic representations and map new information onto the developing structure, the first-mentioned item provides a pathway for the sentence-level integration of incoming information later, rendering that item advantageous and privileged in comprehension (Gernsbacher, 1990). Moreover, this strategy aligns with the typical composition of an event by placing an entity that engages most strongly with an action in the early phase of information flow (Bornkessel-Schlesewsky & Schlesewsky, 2009; Cohn & Paczynski, 2013).

Because of its motivation from multiple sources, this strategy is often deemed as the interface of linguistic knowledge and domain-general factors in the human mind (Bever, 1970; Bornkessel-Schlesewsky & Schlesewsky, 2009; Esaulova et al., 2021; Ferreira, 2003; Givón, 1995; Kemmerer, 2012). Indeed, this strategy has drawn attention to researchers working on child language; existing literature, mostly based on a limited range of languages, reports children's heavy reliance on this strategy for sentence comprehension (Abbot-Smith et al., 2017; Cristante & Schimke, 2020; Gertner et al., 2006; Jackendoff & Wittenberg, 2014; Sinclair & Bronckart, 1972; Slobin & Bever, 1982; Yuan et al., 2012). This favours the early emergence and universal application of this strategy as an intrinsic cognitive bias for child comprehension across languages (but see Garcia & Kidd, 2020; Shin, 2021). Several studies have modelled word-order preferences (broadly touching upon this strategy) in production at a satisfactory level (Chang et al., 2006), together with cross-linguistic variability (Chang, 2009). What remains is to see if this success also holds for comprehension, a process in which a language user identifies an intended meaning/function from the given form (Goldberg, 2019), without positing pre-determined/artificial sets of input (e.g., form-meaning pairs) for model composition as the previous studies did.

1.1 | *Agent-First* strategy in comprehension of Korean active transitive

We pursue this inquiry through an active transitive construction in Korean, an agglutinative, Subject–Object–Verb language with overt case-marking and understudied for this topic. The canonical word order for the active transitive follows agent–theme ordering (1a); this can be scrambled (1b), manifesting the reverse thematic role ordering (theme–agent). Korean allows the omission of sentential components if the omitted information can be inferred from the context (Sohn, 1999). As long as participants in an event are clearly identified in the context, a case marker (2a) or a combination of an argument and a case marker (2b) can be omitted without changing the basic propositional meaning.

(1a) Active transitive (canonical)

kyengchal-i	totwuk-ul	cap-ass-ta.
police-NOM	thief-ACC	catch-PST-SE[1]

'The police caught the thief.'

(1b) Active transitive (scrambled)

totwuk-ul	kyengchal-i	cap-ass-ta.
thief-ACC	police-NOM	catch-PST-SE

'The police caught the thief.'

(2a) Omission (case marker)

kyengchal-i	totwuk-~~ul~~	cap-ass-ta.
police-NOM	thief-~~ACC~~	catch-PST-SE

'The police caught the thief.'

(2b) Omission (case-marked argument)

~~kyengchal-i~~	totwuk-ul	cap-ass-ta.
~~police-NOM~~	thief-ACC	catch-PST-SE

'The police caught the thief.'

Previous literature on Korean-speaking children's comprehension has reported that the canonical pattern is more reliably interpreted than the scrambled one, with the sentence-initial argument mapped

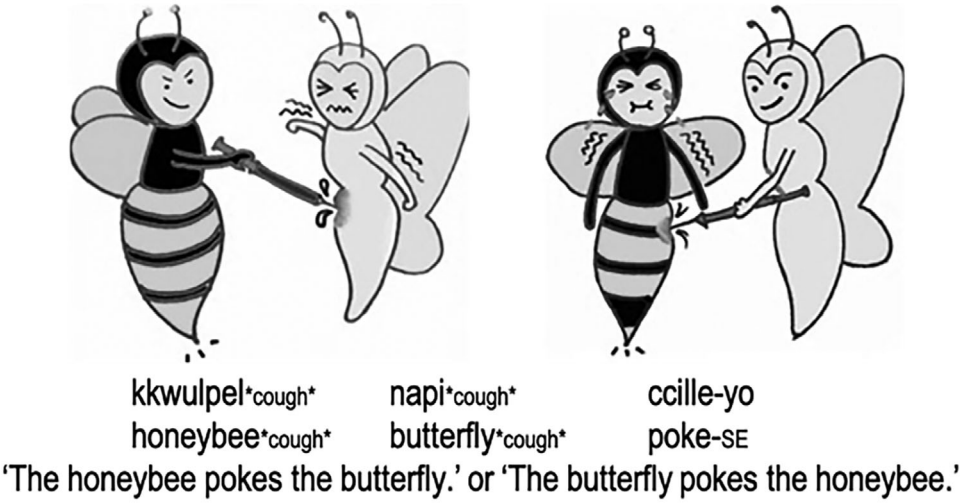


FIGURE 1 Example of test stimuli used in Shin (2021): $N_{CASE}N_{CASE}V$.

onto the agent (regardless of its actual thematic role) until the age of four (Cho, 1982; Kim et al., 2017). This indicates the children's utilisation of the *Agent-First* strategy as the default comprehension bias, as found in many languages.

However, Shin (2021) reveals some limits on this bias, arguing that there may be no standalone *Agent-First* strategy for comprehension (see also Garcia & Kidd, 2020). Shin finds that, for Korean-speaking children's comprehension of a transitive event, the *Agent-First* strategy is activated properly only in conjunction with other types of grammatical cues. Shin measured typically developing 3–6-year-old children's comprehension of the active transitive construction involving scrambling/omission of constructional components through picture-selection tasks with an innovative methodology that systematically obscured parts of test stimuli with acoustic masking (Figure 1).

Shin (2021) notes four major findings (Table 1). First, whereas the children had a good command of case-marking knowledge regarding the active transitive (the nominative case marker indicating the agent; the accusative case marker indicating the theme), they showed an asymmetry in performance by canonicity: they were better in the canonical condition ($N_{NOM}N_{ACC}V$) than in the scrambled condition ($N_{ACC}N_{NOM}V$). Second, they did not manifest the agent-first interpretation strongly in $N_{CASE}V$, showing around 40% for the 3-and-4-year-olds and around 60% for the 5-and-6-year-olds (and 67% at best for the adult controls). In this condition, children must determine the thematic role of the first and the sole case-less argument, which can in principle be interpreted as either the agent or the theme. If the *Agent-First* strategy strongly guides children's comprehension, this argument should be interpreted as the agent reliably, which was not the case. Third, compared to $N_{CASE}V$, the presence of a second noun ($N_{CASE}N_{CASE}V$) increased responses consistent with the *Agent-First* strategy, but its magnitude differed by age such that only the 3-and-4-year-olds considerably enhanced the agent-first interpretation from $N_{CASE}V$ to $N_{CASE}N_{CASE}V$. Fourth, the presence of case markers substantially increased the agent-first response rates for both age groups, as shown in $N_{NOM}V$.

Based on these findings, Shin (2021) argues that, when Korean-speaking children interpret a transitive event, they do not employ this strategy automatically and immediately based solely on an argument's initial position in the sentence. Considering the particular experimental setting in which participants were exposed to pictures prior to stimuli so that they adjust their interpretation to transitive events with two animate entities (one as an agent and the other as a theme) before encountering the stimuli, the children's comprehension behaviour would have been guided by two major forces. One involves properties of caregiver input regarding transitive events. In CHILDES, the number of first-noun-as-agent pattern instances did not exceed that of first-noun-as-theme pattern instances, but almost all of the transitive instances had either a second argument or a marker (with a strong association between the agent and the nominative case marker). The other force involves the developing nature of a child processor, prioritising a local cue over a distributional cue (Wittek & Tomasello, 2005) when dealing with various (non-)grammatical cues simultaneously to accomplish the task at hand. Children may thus attend to the local pairing that associates the nominative-marked argument onto agenthood before becoming sensitive to the broad-scope distributional cue involving a second argument in employing the assumed *Agent-First* strategy for a complete interpretation of a transitive sentence at hand. Because the activation of the *Agent-First* strategy is tied to other grammatical cues such as case-marking (as a local cue; particularly the nominative case marker) and a second nominal (as a distributional cue), Korean-speaking children (and even adults) employ this strategy with confidence only when they are provided with a linguistically informative environment. This argument challenges the long-standing idea that children have the default mapping of the agent onto the first noun as an intrinsic bias for comprehension, as claimed by previous studies targeting the major languages being investigated (Abbot-Smith et al., 2017; Cristante & Schimke, 2020; Gertner et al., 2006).

TABLE 1 Summary of results: major conditions.

Condition	Group	Mean (%)	SD	Note
$N_{NOM}N_{ACC}V$	3–4-year-olds	84.44	0.36	Scoring: accuracy (1: correct; 0: incorrect)
	5–6-year-olds	94.20	0.24	
	Adult	100.00	0.00	
$N_{ACC}N_{NOM}V$	3–4-year-olds	77.78	0.42	
	5–6-year-olds	71.01	0.46	
	Adult	100.00	0.00	
$N_{NOM}V$	3–4-year-olds	94.44	0.23	
	5–6-year-olds	97.10	0.17	
	Adult	93.33	0.25	
$N_{ACC}V$	3–4-year-olds	92.22	0.27	
	5–6-year-olds	97.10	0.17	
	Adult	100.00	0.00	
$N_{CASE}N_{CASE}V$	3–4-year-olds	66.67	0.48	Scoring: high likelihood of *agent-first* interpretation (1: agent-first; 0: theme-first)
	5–6-year-olds	77.27	0.42	
	Adult	90.00	0.04	
$N_{CASE}V$	3–4-year-olds	42.59	0.50	
	5–6-year-olds	60.42	0.49	
	Adult	66.67	0.06	

Abbreviations: ACC, accusative case marker; CASE, case marker (unspecified); NOM, nominative case marker.

1.2 | The present study

We investigate whether and how NNs, as a proxy for cognitive space wherein learning occurs, reveal children's manifestation of the *Agent-First* strategy in comprehension. We develop four NN models—*Word2Vec* (Mikolov et al., 2013), Long Short-Term Memory (*LSTM*; Hochreiter & Schmidhuber, 1997), Bidirectional Encoder Representations from Transformers (*BERT*; Devlin et al., 2018), Generative Pre-trained Transformer 2 (*GPT-2*; Radford et al., 2019)—and measure their classification performance on the same stimuli used in Shin (2021). Given the special status of this strategy in child language development as a window to the interface between linguistic knowledge and domain-general factors, scrutinising the extent to which deep-learning algorithms capture children's language behaviour with respect to this comprehension bias is expected to reveal the explainability of artificial intelligence for child language, and more fundamentally, for (the developing nature of) a child processor.

Word2Vec is a two-layer NN algorithm that creates word embeddings through information about words given their local usage contexts, by converting each word into multi-dimensional vectors and calculating the similarity between these vectors. Two model architectures comprise distributed representations of words: 'continuous bag-of-words' predicting the current word given its surrounding words (word order does not affect this process); 'continuous skip-gram' predicting the surrounding words given the current word. *LSTM* is a recurrent NN algorithm which is capable of handling long-term dependencies by allowing information to persist. This architecture is characterised as the hidden layer comprising a memory cell with three gates: Forget (determining whether the incoming information from the previous timestamp is irrelevant and thus forgotten; Input (quantifying the significance of new information carried by the incoming input); Output (submitting the currently updated information to the next timestamp). *BERT* and *GPT-2* share the transformer architecture, utilising the attention mechanism for effective computation. This mechanism enhances each part of the input sequence differently, considering various information about the whole sequence (e.g., segment position), to better identify the most relevant parts of that sequence (Vaswani et al., 2017). This enhancement allows the transformer to retain information from the early-appearing elements when handling long input sequences during information processing (Ludwig et al., 2021; Vaswani et al., 2017). *BERT* obtains rich contextual embeddings through two tasks: masked language model (randomly masking some words in a sentence and predicting these masked words from the context of the exposed words surrounding the masked words); next sentence prediction (determining whether one sentence in a pair would come before or after the other). In contrast, *GPT* targets a general-purpose learner whose learning trajectories are not subject to particular tasks, so model training does not stand on the specifics of data or tasks at hand (Radford et al., 2019); it can also perform new tasks with a relatively small number of examples. Architecture-wise, whereas BERT uses the encoder part and operates non-autoregressively, GPT uses the decoder part and operates autoregressively; other than that, even though technical differences exist in hyperparameters, there is no other notable conceptual difference.

These models have been increasingly used to capture various adult-language features (Futrell & Levy, 2019; Hawkins et al., 2020; Warstadt & Bowman, 2020; Warstadt et al., 2019; Wilcox et al., 2018). You et al. (2021), one recent study relevant to the current work, showed that a Word2Vec learner was able to conduct semantic inference from raw caregiver input without the mediation of structural information, by measuring the model's discrimination performance on English causatives. However, other than You et al. (2021), computational research on child language through NNs is extremely thin.

In the present study, we train each NN model by patching caregiver-input data in CHILDES (MacWhinney, 2000) onto the respective pre-trained models. Caregiver input—which notably differ from adult language usage in terms of clausal composition (e.g., non-human agents, partial utterances) and mode of delivery (e.g., simple, short, repetitive) (Cameron-Faulkner et al., 2003; Shin, 2022a; Stoll et al., 2009)—is known to effectively support children's development of linguistic knowledge (Behrens, 2006; Choi, 1999; Snow, 1972). If NNs faithfully exploit this characteristic for their learning, the models in this study should approximate the children's response patterns measured by Shin (2021), showing reasonable accuracy, like their successful performance in some adult language features (Hawkins et al., 2020; Marvin & Linzen, 2019; Warstadt & Bowman, 2020; Warstadt et al., 2019). Considering the transformer's better capability to capture adult language features than the recurrent architecture (e.g., verb bias in Hawkins et al., 2020), BERT and GPT-2 will be presumably closer than the other models in classification performance relative to children's performance found in Shin (2021). Furthermore, the notable model-specificity involving the two transformers, such as the sole use of the encoder/decoder part in model composition and the GPT algorithm's task-independent nature in the course of model training, would generate differences in classification performance across the two transformers.

2 | METHODS

The general modelling procedure is illustrated in Figure 2.

The caregiver-input data (Appendix A) were pre-processed in two ways: typos and spacing errors were corrected, and any sentence whose length was less than five characters or those consisting only of onomatopoeia and mimetic words were excluded (see Shin, 2022a for the details about the pre-processing). This resulted in 69,498 sentences (285,350 eojeols[2]).

Table 2 summarises the composition of each NN model advised/recommended by previous studies (Church, 2017; Clark et al., 2019; Goldberg & Levy, 2014; Vázquez et al., 2020; Wu et al., 2019).[3] While NNs typically require large-scale data for training to ensure their optimal operation (Edwards, 2015), there is no pre-trained model exclusively constructed with caregiver input, nor a sufficient amount of Korean caregiver input data to create a pre-trained model. In addition, children are not surrounded only with caregiver input in real life; there are many types of exposure to language usage that children experience. To cope with these issues, we employed the respective pre-trained models, which were open-access and representative at the moment of study, in developing each NN model. We believe that adopting a pre-trained model in conjunction with the caregiver-input data can be one way to ensure better ecological validity for the simulation, but apparently, no research has ever touched upon this point, thus worthy of further attention.

For the binary classification of test items (*Agent-First*; *Theme-First*), these models were further trained with instances of all the constructional patterns expressing a transitive event—active transitive and suffixal passive, with scrambling and varying degrees of omission manifested—with labels indicating whether the thematic-role ordering of these instances followed agent-first or theme-first (Appendix B). The instances were extracted from the pre-processed caregiver-input data through an automatic search process developed by Shin (2022a); every sentence for each extraction was also checked manually to ensure its accuracy. Although the focus concerning the *Agent-First* strategy in this study was the active transitive, we included the suffixal passive, another major clause-level device expressing a transitive event and the representative type of passive that children are likely to encounter in caregiver input (Shin & Deen, 2023). Furthermore, considering the zero occurrence of some patterns in the input, we adapted Laplace

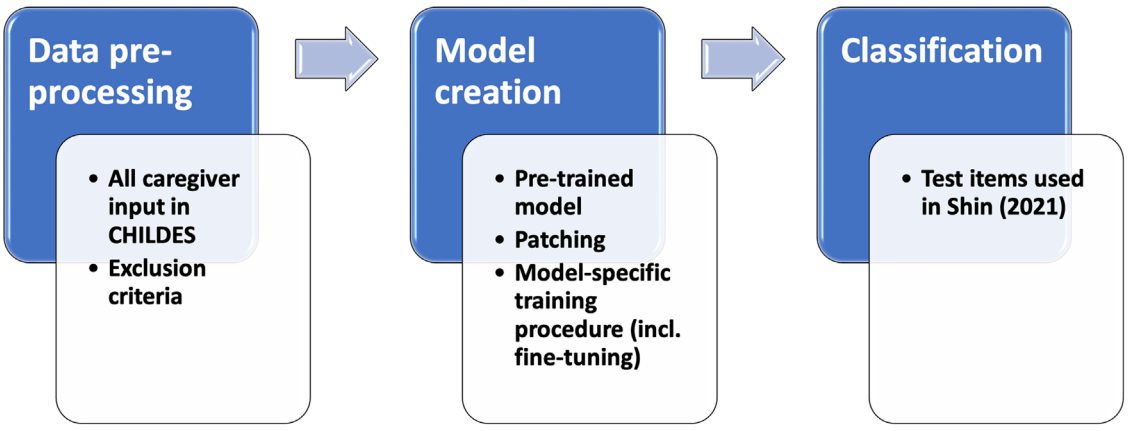


FIGURE 2 General modelling procedure.

TABLE 2 Summary: NN models

	Word2Vec	LSTM	BERT	GPT-2
Python package	***Sklearn, Gensim***	***PyTorch***	***Transformers***	***Transformers***
Pre-trained model	*Pre-trained word vector*[a] [corpus size: 339MB; vocabulary size: 30,185]	*KoCharElectra-Base*[b] [corpus size: undescribed; vocabulary size: 11,568]	*KoBERT*[c] [corpus size: 54MB; vocabulary size: 8002]	*KoGPT2-base-v2*[d] [corpus size: 40GB; vocabulary size: 51,200]
Tokenisation	Morpheme-based	Syllable-based	Syllable-based	Syllable-based
			WordPiece	*Byte Pair Encoding*
Model-specific	Context window fixed (five) as defined by the pre-trained model; classification algorithm: SVM	Hidden layers: 256, epoch: 10, learning rate: 0.00002	Batch: 32, seed: 42, epoch: 10, sequence length: 256, epsilon: 0.00000001, learning rate: 0.0001	

Abbreviations: LSTM, Long Short-Term Memory; BERT, Bidirectional Encoder Representations from Transformers; GPT-2,Generative Pre-trained Transformer 2.

[a] https://github.com/Kyubyong/wordvectors (accessed on 03 November 2021).

[b] https://github.com/monologg/KoCharELECTRA/blob/master/vocab.txt (accessed on 12 October 2021).

[c] https://github.com/SKTBrain/KoBERT (accessed on 15 September 2021).

[d] https://github.com/SKT-AI/KoGPT2 (accessed on 15 September 2021).

smoothing (Agresti & Coull, 1998) by adding one fake instance (following the pattern-wise characteristics) to all the patterns. Nonetheless, most of the input comprised the active transitive, occupying more than 90% of the entire data.

To develop the Word2Vec model, we employed skip-gram negative sampling due to its superior performance in language tasks compared to the 'continuous bag-of-words' approach (Mikolov et al., 2013). For model training, we first patched the caregiver-input data to the pre-trained model, resulting in the change of the total number of word vectors in the pre-trained model (30,185 to 30,638). We then added to the model the transitive-event-related instances as epoch (i.e., a cycle that trains a model with the entire dataset) proceeded up to 10. To conduct the classification task, we first separated sentences from labels in the caregiver-input data, tokenised all the sentences by morpheme, and created a new word-embedding model with morpheme vectors given the pre-trained model (which was morpheme-based with a fixed context-window size). We then reduced the dimension of the new model down to one by using t-SNE (Maaten & Hinton, 2008). Word2Vec generates word vectors but does not perform classification, so we employed support vector machine (SVM; Cortes & Vapnik, 1995) for the planned task (cf. Abdelwahab & Elmaghraby, 2016). The new embedding model, together with the label information per sentence, applied to train the SVM classifier by converting each sentence's morpheme in the input to either one (when that morpheme existed in the model) or zero (when that morpheme did not exist in the model). The trained classifier ultimately predicted if the label of a test stimulus was *Agent-First* or *Theme-First*. For this model-classifier combination, no variation per trial in each epoch occurred due to the invariant nature of the word embeddings generated by Word2Vec.

In developing the LSTM model (with syllable-based tokenisation; Table 2), there exists no syllable-based Korean pre-trained model for LSTM, so we adapted a pre-trained model for ELECTRA to extract relevant vocabulary information to train the model. After patching the caregiver-input data to the pre-trained model, we found no change in the model size, meaning that all the syllable types in the caregiver-input data were already included in the pre-trained model. We then added to the model the transitive-event-related instances as epoch proceeded up to 10. For each epoch, all the syllable information was submitted to the model's input layer. Take an eojeol *saca-ka* 'lion-NOM' as an example (see Figure 3 for illustration). For the syllable *ca*, the model first evaluates if the information about the previous syllable *sa* obtained from the prior cell is relevant to the current input at the Forget gate (σ_1). The model then quantifies the information about the current input via the tangent function at the Input gate (σ_2). Finally, the model hands over this outcome to the processing of the next syllable *ka* at the Output gate (σ_3), again via the tangent function. Once a sentence is complete for processing, the optimiser computes the distance/loss between the observed value and the predicted value, the result of which is transmitted through backpropagation. In our model, the loss value was reset after 500 sentences. Once the training was completed, the model evaluated the test stimuli, accumulating by-syllable information sequentially (by generating respective hidden layers) and then comparing the outcomes (1 = *Agent-First*; 0 = *Theme-First*) to the actual

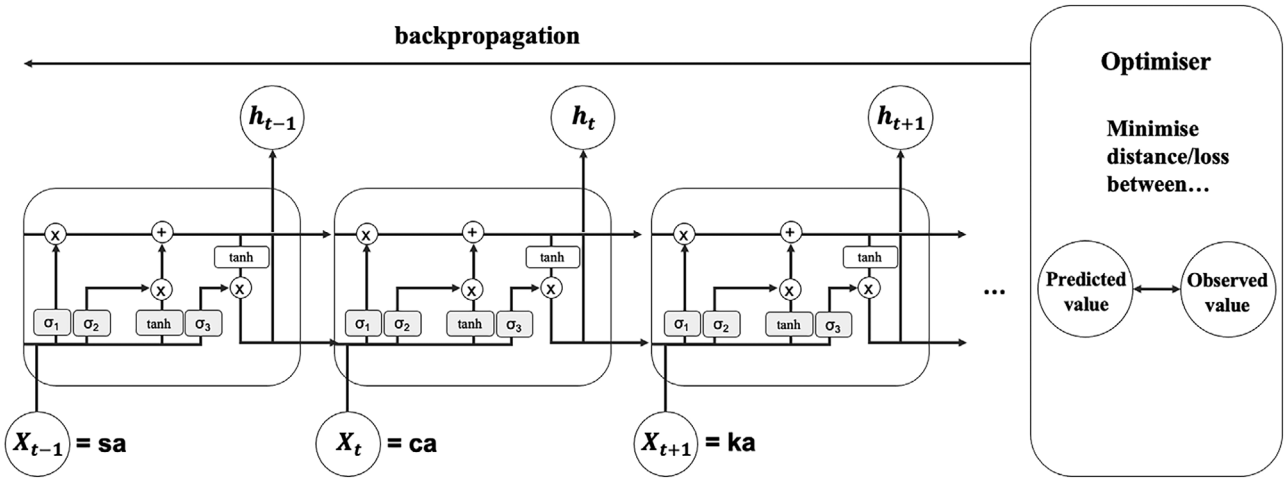


FIGURE 3 Model training: LSTM (e.g., *saca-ka* 'lion-NOM').

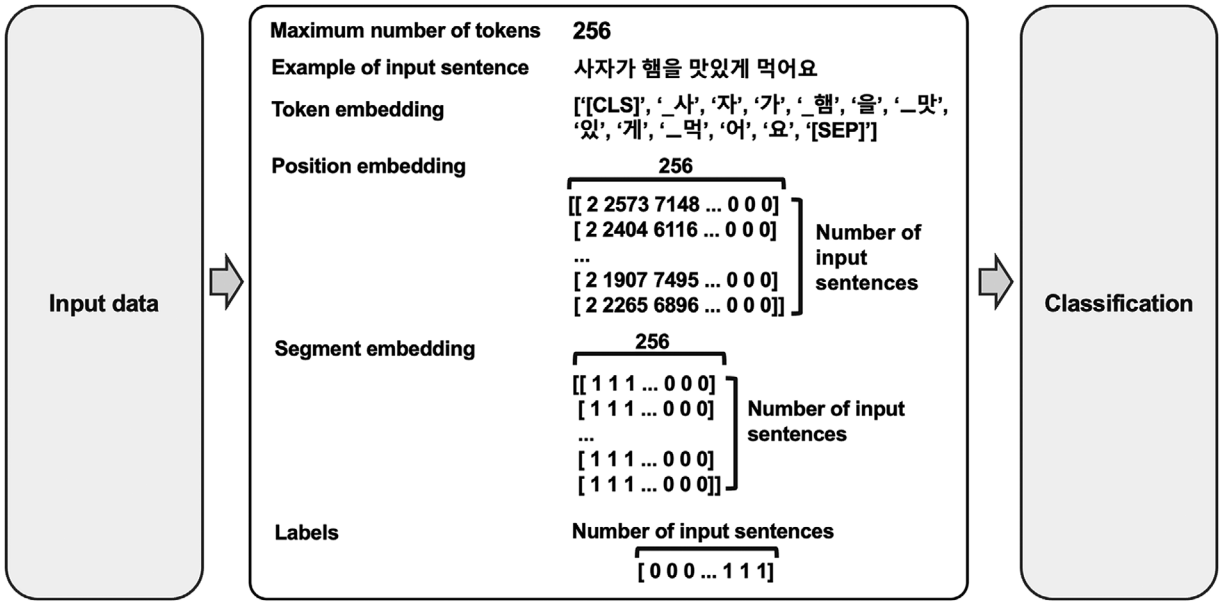


FIGURE 4 Model training: BERT (e.g., *saca-ka haym-ul masiss-key mek-eyo* 'lion-NOM ham-ACC delicious-ADV eat-SE' "The lion eats the ham deliciously").

labels of these stimuli. We repeated the same learning process 30 times in each epoch and averaged the by-condition outcomes in assessing the models' classification performance to alleviate potential variations during the task.

For the BERT model (see Figure 4 for model fine-tuning), every input sentence began and ended with [CLS] (marking the start of a sentence) and [SEP] (marking the end of a sentence) to indicate sentence boundaries, and the length of each sentence was limited to 256 tokens. The patching procedure increased the pre-trained model's vocabulary size (8002 to 24,857). We then added to the model the transitive-event-related instances as epoch proceeded up to 10, with relevant labels (*Agent-First*; *Theme-First*) attached. Each input sentence in the fine-tuning stage was transformed into three embedding types. For token embedding (see Figure 4 for illustration), the sentences were tokenised (as a unit of syllable). For position embedding, each token was converted into a numeric value indicating a unique index of the token with reference to the vocabulary in the patched pre-trained model (*KoBERT*). For segment embedding, the numeric value of 1 was mapped onto a slot when a token occurred in that position of a sentence; otherwise, the numeric value of 0 was used. The initial values of epsilon (i.e., the upper bound of randomness for a model to explore the data), learning rate (i.e., the degree to which a model changes in response to the estimated error when the model weights are updated), and seed (i.e., the initialisation state of a pseudo-random number generator indicating where a model starts) were automatically updated with the outcomes of each epoch. The training occurred 320 times (32 batches

TABLE 3 Composition of test stimuli

Condition	Example	Expected classification
$N_{NOM}N_{ACC}V$	Honeybee-NOM butterfly-ACC poke	Agent-first
$N_{ACC}N_{NOM}V$	Butterfly-ACC honeybee-NOM poke	Theme-first
$N_{NOM}V$	Honeybee-NOM poke	Agent-first
$N_{ACC}V$	Butterfly-ACC poke	Theme-first
$N_{CASE}N_{CASE}V$	Honeybee butterfly poke	Agent-first
$N_{CASE}V$	Honeybee poke	Agent-first

Abbreviations: ACC, accusative case marker; CASE, case marker (unspecified); NOM, nominative case marker.

[the number of samples—rows of data—passing through to a model at one time] * 10 epochs) from the initial model with the zero value of gradients to an optimal model with updated values through feedforward and backpropagation (cf. Xu et al., 2020). Finally, the trained model per epoch classified the test stimuli; likewise for the LSTM model, we averaged the by-condition classification outcomes from 30 times of learning.

The GPT-2 model's training process was almost the same as above, except that GPT uses no symbol to mark the start/end of each input sentence; after patching the caregiver-input data to the pre-trained model, its vocabulary size increased (51,200 to 67,052). Originally, GPT and BERT differ with respect to tokenisation: while BERT (*WordPiece*) utilises a word as a basis for tokenisation, GPT-2 (*Byte Pair Encoding*) utilises a character (in the case of English) for this purpose. However, both *KoBERT* and *KoGPT-2* employ a syllable as a basic unit of tokenisation (likely in consideration of the properties of Korean), so there was no essential difference between the two methods regarding tokenisation (but note that the two models manifest notable model specificity; see Section 1.2).

For test items, we employed the same stimuli used in Shin (2021). Each condition consisted of six instances, with animals as agents and themes and actional verbs at the end (Table 3). Each trained model classified every test stimulus, evaluating whether the stimulus fell into Agent-First or Theme-First. We note that, while the stimuli of $N_{CASE}N_{CASE}V$ and $N_{CASE}V$ in Shin (2021) involved acoustic masking, the same stimuli type in the simulation did not have such auditory effects. This was unavoidable considering this study's simulation setting where the models worked exclusively with the text data. We acknowledge that this difference might serve as one confounding factor for interpreting the results.

3 | RESULTS

3.1 | Case-marked conditions

Figure 5 illustrates the classification performance of the four models on the four case-marked conditions. For the two-argument conditions, whereas all the models except BERT achieved high accuracy in $N_{NOM}N_{ACC}V$, only LSTM demonstrated high accuracy in $N_{ACC}N_{NOM}V$. For the one-argument conditions, all the models except BERT showed high accuracy in $N_{NOM}V$ and the four models exhibited very high accuracy in $N_{ACC}V$. Overall, of the four models, LSTM seemed close to children in its performance on these conditions.

While the BERT model's performance appeared to be peculiar, the results broadly imply two possible traits in the models' classification performance on the case-marked conditions. First, it seems that Word2Vec and GPT-2 followed characteristics of the caregiver input selectively. There are two important characteristics of the caregiver input (Appendix B). One is that the number of first-noun-as-agent patterns (3049 instances) did not exceed that of first-noun-as-theme patterns (3579 instances). The other property is that the number of nominative-first patterns (overtly marked with the nominative case marker; 3369 instances) outnumbered that of accusative-first patterns (overtly marked with the accusative case marker; 1989 instances) despite the generally higher omission rate of the accusative case marker than that of the nominative case marker in caregiver input (Shin, 2022a). Given these characteristics, as epoch progressed, the two models may have attended primarily to the form of a specific case marker (overtly attested in a test stimulus) rather than to the meaning/function (i.e., thematic roles) of the initial noun, possibly leading to both success in one-argument conditions where consideration of thematic role ordering was not required but partial success in the two-argument conditions where thematic role ordering between the two arguments should be considered. This may have been further enhanced by the respective pre-trained models, created by general/adult language use involving the dominance of canonical word order and the frequent omission of the accusative case marker (Sohn, 1999).

Second, the LSTM model's outperformance over the transformers—against our prediction—possibly indicates the algorithm-exclusive memory cell's contribution to information processing. That is, the existence of a memory cell may have assisted the classification accuracy as effectively as the attention mechanism of the transformers in the given simulation environment. Considering that transformer architecture excels in utilising information from long input sequences (see Section 3), it is reasonable to think that the transformers in this study (and BERT in particular) may not have fully exerted their algorithmic strength when coping with child language behaviour. The LSTM model's good classification performance further aligns with previous reports on the LSTM's success in learning and generalising clause-level linguistic knowledge (Futrell & Levy, 2019; Marvin & Linzen, 2019; Wilcox et al., 2018). Specifically, when the characteristics of a test stimulus does not match those of typically appearing sentences in use (like scrambled word order), the attention mechanism may not have discriminated that stimulus effectively due to the larger volume of information—both sequential and positional information—that it retains compared to the recurrent architecture, which has only sequential information. This implies that a sophisticated, cutting-edge model may not always bring the best outcome.

However, these interpretations only tentatively answer how these NN models reveal children's utilisation of the *Agent-First* strategy in comprehension, until we check the models' performance in the

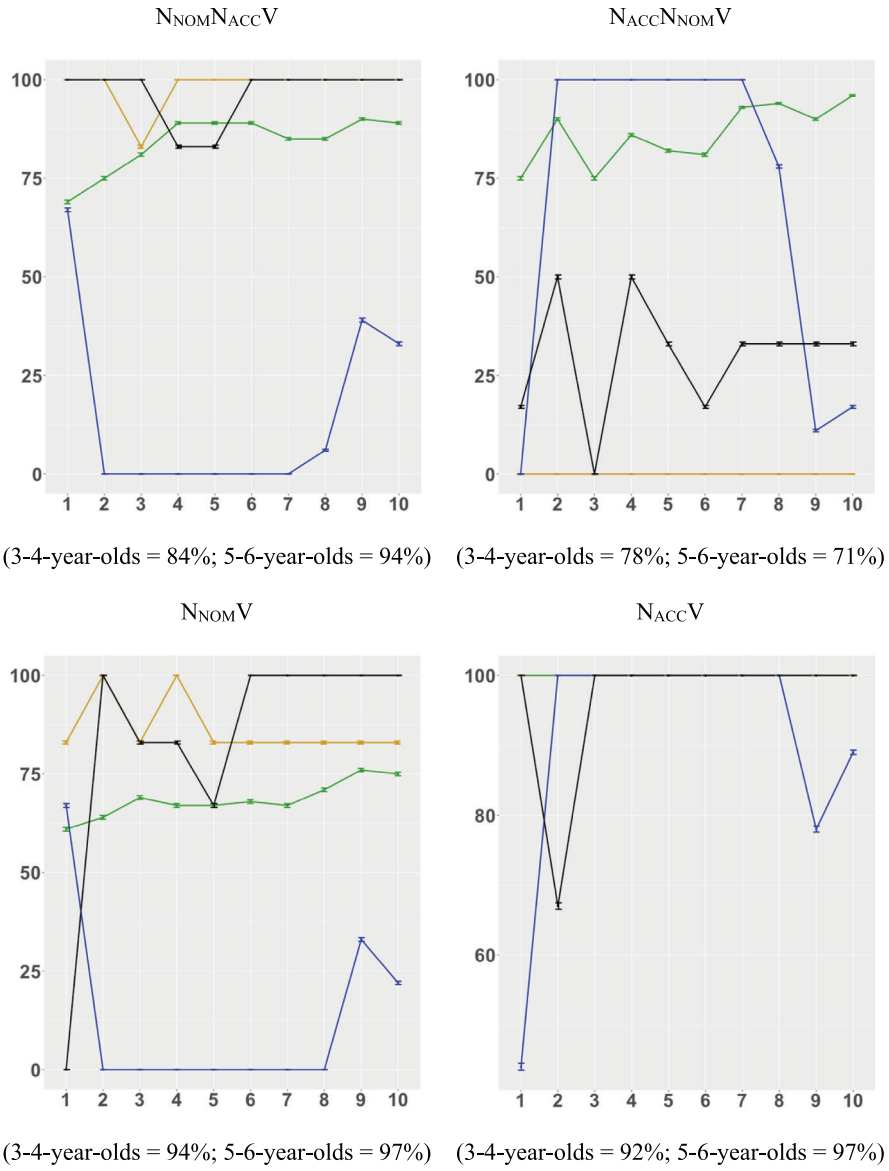


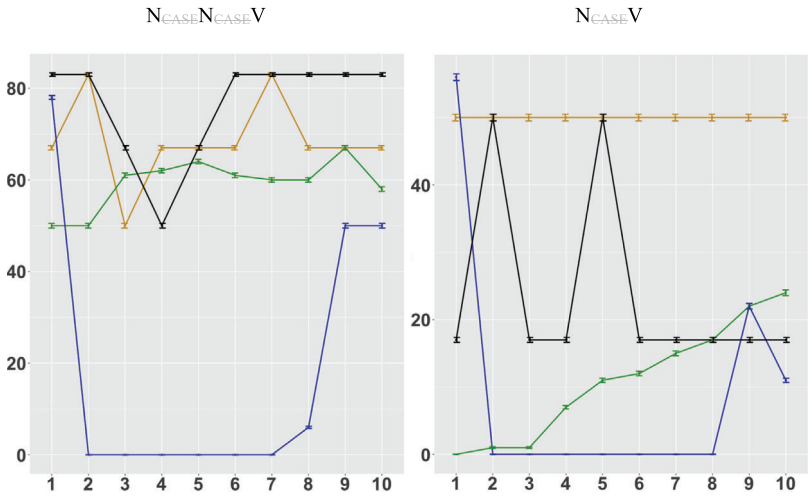
FIGURE 5 Child comprehension and model performance: case-marked conditions. *Note*: x-axis: epoch; y-axis: accuracy (averaged). Gold = Word2Vec; Green = LSTM; Blue = BERT; Black = GPT-2. Error bars indicate 95% CI.

remaining two case-less conditions—$N_{CASE}N_{CASE}V$ and $N_{CASE}V$. The next section presents the models' performance on these two conditions, which had served as the core evidence against the children's spontaneous and faithful application of this cognitive bias towards comprehension in Shin (2021).

3.2 | Case-less conditions

Figure 6 illustrates the classification performance of the four models on the two case-less conditions. For $N_{CASE}N_{CASE}V$, all the models except BERT showed an agent-first preference in classification as epoch progressed (with GPT-2 being the highest), which was broadly similar to the children's performance in this condition. Notably, BERT underperformed in this condition, with the chance-level rate of agent-first classification at epoch 10. For the $N_{CASE}V$, all the models demonstrated eccentric performance: Word2Vec invariably remained at-chance; LSTM steadily improved its performance (but its agent-first classification rate was under 30%); the performance of BERT and GPT-2 fluctuated considerably and these models yielded less than 20% of agent-first classification rate.

Overall, in these case-less conditions, the NN models failed to capture the trend manifested by the children in a satisfactory manner. Specifically, the two transformers (BERT; GPT-2) malfunctioned in $N_{CASE}V$, performing with high deviation from the children's interpretation for the same condition. One possible cause of this global anomaly originates from the interaction between the nature of the two conditions and the models' information-processing mechanism, which looks exclusively to formal sequences. Recall that the two conditions involve no case-marking; this under-informativeness in determining



(3-4-year-olds = 67%; 5-6-year-olds = 77%) (3-4-year-olds = 43%; 5-6-year-olds = 60%)

FIGURE 6 Child comprehension and model performance: case-less conditions. *Note*: x-axis: epoch; y-axis: agent-first classification rate (averaged). Gold = Word2Vec; Green = LSTM; Blue = BERT; Black = GPT-2. Error bars indicate 95% CI. The performance indicates the high likelihood of agent-first interpretation (1: agent-first; 0: theme-first) because these conditions can in principle be interpreted in more than one way.

the thematic role of the first noun would have affected both the children's comprehension and these models' classification performance. However, the NN models may have been more influenced than the children by the lack of reference point for the classification decision (i.e., case marker) that was attested in the stimuli, rendering their performance substantially deviant from the children's response rates. The fact that the LSTM model improved its classification performance towards Agent-First as epoch progressed may further indicate the contribution of its algorithmic characteristic to revealing child language behaviour as in the case-marked conditions, but notably, its agent-first classification rate was still far below the children's performance.

In sum, compared to the children's interpretation, the classification performance of the NN models on the two case-less conditions was altogether eccentric, manifesting asymmetric degrees of by-model and by-condition performance.

4 | GENERAL DISCUSSION

Motivated by the proxy provided by NNs as biologically inspired models of computation, we developed four NN models and measured their classification performance on the active transitive sentences used in Shin (2021) involving scrambling/omission of sentential components at varying degrees. Considering Shin's (2021) experimental setting in which the children were shown transitive-event pictures prior to a stimulus to contextualise their interpretation of that stimulus, we trained each model with the caregiver input of constructional patterns for expressing transitive events in CHILDES. Overall, despite some compatibility of these models' performance with the children's response patterns, their performance did not fully approximate the Korean-speaking children's comprehension behaviour pertaining to the *Agent-First* strategy, and demonstrated by-model and by-condition asymmetries. Moreover, the predicted benefit of transformers in this classification task did not clearly emerge.

This study's results are ascribable to various factors. For instance, the simulation environment in this study may not have sufficiently conformed to Shin's (2021) experimental setting to the extent that the models processed the stimuli in the same way as the children did in the experiment. Recall that we trained each model with all the transitive-event instances in CHILDES (see Appendix B), reflecting how the children in Shin (2021) attuned their interpretation to transitive events before they were exposed to the stimuli. Despite this treatment, the models might not have had a testing environment fully compatible with what the children experienced. Moreover, the test items in the simulation involved no acoustic signals (Table 3) as used in Shin (2021) that allowed the children to know that there was something but hidden (Figure 1). This absence of auditory information about the marker(s), which was inevitable given the simulation setting in which the models operated exclusively with the textual data, may thus have affected the model performance in an unexpected way (cf. Stoyneshka et al., 2010).[4] On top of this, we used the pre-trained models involving mature language in various genres when constructing each NN model for various reasons (see Section 2). Together, although we conducted the simulation work as consistently with the experimental setting in Shin (2021) as possible, this simulation inherently stood on a slightly different ground than the experiment (as most modelling research does), possibly generating the observed model-children asymmetry. However, we highlight that, because these issues have not been fully explored in this field, we cannot conclude that these are the all-and-only reasons of this asymmetry.

Another possible contributing factor to the model performance is around language-specific properties. Whereas Korean caregiver input joins the general characteristics of child-directed speech (Shin, 2022a; cf. Cameron-Faulkner et al., 2003; Stoll et al., 2009), it also manifests language-specific properties such as scrambling and omission of sentential components (see Appendix B for the constructional-pattern-wise variability). Along with the general nature of caregiver input, the models may thus have been affected by the specific word order and/or the presence of case markers in conducting the classification,

particularly as shown with the two-argument case-marked scrambled condition ($N_{ACC}N_{NOM}V$) and the case-less conditions ($N_{CASE}N_{CASE}V$; $N_{CASE}V$). This aligns with previous reports on language-specific challenges for automatic processing of Korean (Kim et al., 2007; Shin, 2022a). Since we are not aware of any study on language-specific properties and NNs' performance on child language, this claim awaits further examination.

In addition to these factors, we argue that the models' algorithmic characteristics may be a core source of this asymmetry. NNs often exploit contextual information through window-based computation (Haykin, 2009; Kriesel, 2007) when given a sampling of data points. One common practice regarding this computation is to induce contextual information from formal sequences comprising words and/or characters; to put it differently, they rely heavily on form. This yields a context in a computational sense (cf. Firth, 1957), but it differs from a context in a linguistic sense comprising semantic–pragmatic information. Hence, whenever the models access the meaning/function of a linguistic unit, they exploit the formal co-occurrences in the incoming input, rather than directly drawing upon the meaning/function of that unit. Moreover, NNs are designed to generalise what they already have (through pre-trained models and additional information obtained from training) but are not designed to make reasonable predictions and extrapolation outside of the training space (Marcus, 1998). Deep-learning models attempt to resolve this issue by using exceedingly large datasets to cover every possible instance of formal co-occurrences; this often yields good performance when handling known inputs, but still not with novel inputs.[5] The stimuli in Shin (2021), consisting of animal names as entities, would be new instances for these NN models in this respect (and also considering the typical composition of transitive sentences in ordinary speech—animate agents and inanimate themes; e.g., Dowty, 1991; Ibbotson & Tomasello, 2009; Langacker, 1991), thus possibly leading the models to malfunction in their operation. Therefore, this algorithmic nature may have rendered the NN models deviant from the children's performance on some test items possibly out of range. The key evidence comes from the models' performance on $N_{CASE}V$ (where a simulated learner must determine the thematic role of the first and sole case-less noun only with its presence) compared to their performance on $N_{NOM}V$ and $N_{ACC}V$ (where a simulated learner has more, and core, information about the first noun's thematic role indicated by specific case marker next to the noun).

Relating to this, the reason for the peculiarity of the BERT model in this study is unclear. We speculate that the unit of processing (i.e., sentence) and the way that it learns through two particular tasks may not be ideal for processing the test items given the notable differences between the experimental setting and the simulation environment. Considering that BERT often demonstrates good performance with long sequences (Devlin et al., 2018; Vaswani et al., 2017), the simple, short, and repetitive nature of child-directed speech may have diluted its algorithmic strength. Compared to that, the domain-generality of the GPT algorithm may have offset the similar drawback to some degree, leading to a partial success in approximating the children's response rate (but not to the extent that the LSTM model showed its gradual improvement regarding the performance on $N_{CASE}V$). Regardless, our reasoning here remains speculative and requires further investigation.

Despite the same pursuit of efficiency in information processing, this manner of algorithmic operation differs from how a human processor deals with linguistic knowledge. Decades of research have shown that the linguistic processor operates in a way that reduces the burden of work at hand, by immediately mapping form onto function (and vice versa) under simultaneous activation of multiple (non-)linguistic routes, combined with cognitive-psychological factors (Christianson, 2016; Ferreira, 2003; Karimi & Ferreira, 2016; Levy, 2008; McElree, 2000; McRae & Matsuki, 2009; O'Grady, 2015; Traxler, 2014). In particular, the child processor manifests notable characteristics due to its the developing nature (cf. Omaki & Lidz, 2015). To illustrate, the child processor favours reliable and/or available cues with a one-to-one mapping relation between form and function (Bates & MacWhinney, 1989; Cameron-Faulkner et al., 2003; Shin, 2021, 2022b; Shin & Mun, 2023). Given the global impact of general language-usage experience (Ambridge et al., 2015; Tomasello, 2003), the processor is particularly sensitive to particular linguistic environments in which a target item is situated (Dąbrowska, 2008; Dittmar et al., 2014; Goldberg et al., 2004). The degree to which the current stimulus is informative against the prior language-usage experience also modulates its performance (Dittmar et al., 2008; Shin, 2021, 2022b; Shin & Deen, 2023; Stromswold et al., 1985). Furthermore, the contribution of domain-general factors to the processor's operation is sometimes limited or less efficient (Adams & Gathercole, 2000; Diamond, 1985). These aspects seem to collectively modulate how the developing processor adjusts its way to arriving at comprehension (Choi & Trueswell, 2010; Garcia et al., 2021; Huang et al., 2013; Özge et al., 2019; Snedeker & Trueswell, 2004).

Therefore, it may be the case that the children in Shin (2021) made the best, albeit imperfect, use of the information available at the time, based on their learning trajectories. When the children listened to an aural stimulus and were asked to choose one picture over the other, they must compute the relative agenthood between the two arguments with no animacy cue available (cf. Chan et al., 2009; Theakston et al., 2012). For this task, the child processor was likely to draw upon multiple morpho-syntactic and semantic cues, including distributional (e.g., mapping between an event representation and a syntactic representation manifested in word order) and local (e.g., pairings between thematic roles and case-marking) ones, which are searchable from their language-usage profiles. At the same time, their interpretation was likely to be swayed away by multiple sources, including event/world knowledge, memory operation, and a cognitive bias such as the *Agent-First* strategy. This simultaneous interplay of various linguistic and domain-general factors affecting the operation of the child processor may not have been properly captured/modelled by the NN models developed in the current study.

Notably, the findings of the present study do not entirely align with those of You et al. (2021), which assumed that word-embedding algorithms resemble human distributional learning by building connections between contextual frames and target words. Based on the success

of the Word2Vec learner in the discrimination task, they claimed that the Word2Vec's ability to extract causal meaning from simple co-occurrence of neighbouring words suffices in modelling how a statistical learner acquires the intended meaning from the formal patterns in raw input. While we agree with the role of contextual information (in a computational sense) generated by formal correlations from input sequences for model performance, we hesitate to fully advocate this claim about child language development. As shown in this study, there are limits on the success of a computational model in approximating children's comprehension behaviour only by accessing word co-occurrences in caregiver input.[6] Moreover, You et al.'s study relied on one model (Word2Vec) and one language (English), rendering it difficult to precisely assess the NNs' ability concerning this issue. This calls for further study seeking to clarify what NNs can(not) explain pertaining to child language development.

5 | CONCLUSION

It appears that NNs tested in this study can utilise information about formal co-occurrences to access the intended message to a certain degree. However, (the outcome of) this process may substantially differ from how a child, as a developing processor, engages in comprehension. Through its use of various NN models and language typologically different from the major languages currently under investigation in the field, the present study provides evidence of some limits of the NNs' capability to address child language behaviour, with a special focus on the *Agent-First* strategy in child comprehension. We believe the implications of this study invite subsequent questions of the extent to which NNs, as an artifact of biological neurons, reveal developmental trajectories of child language that have been unveiled through corpus-based and experimental research. This proffers an important avenue for future research on the explainability of artificial intelligence on (child) language development typically surrounded with various linguistic and cognitive-psychological factors and situated under various usage/learning contexts.

ACKNOWLEDGEMENTS

This study was supported by the European Regional Development Fund through the 'Sinophone Borderlands—Interaction at the Edges' project (CZ.02.1.01/0.0/0.0/16_019/0000791) and the National Research Foundation of Korea (NRF-2022S1A5C2A02090368).

CONFLICT OF INTEREST STATEMENT

The authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in CHILDES (MacWhinney, 2000). MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

ORCID

Gyu-Ho Shin https://orcid.org/0000-0002-8157-7148

ENDNOTES

[1] Abbreviations: ACC, accusative case marker; ADV, adverbial suffix; CASE, case marker (unspecified); NOM, nominative case marker; PST, past tense marker; SE, sentence ender; V, verb.

[2] An eojeol, roughly corresponding to a word in English, is defined as a unit with white space on both sides that serves as the minimal unit of sentential components.

[3] See this repository for the code and dataset. On a side note, comparing variations of the same NN architecture with parameter manipulation was not the primary interest in this study; we followed recommendations and suggestions made by previous studies to obtain optimal outcomes. We believe this model-internal comparison would present a robust area for future research.

[4] Some studies try to alleviate this issue by implementing additional devices to their simulations such as thematic role variables (Chang, 2002) and a layer encoding semantic information (Alishahi & Stevenson, 2008).

[5] In this respect, one promising direction of future research would be to consider multimodal embedding in modelling child language (cf. Mithun et al., 2018; Sung et al., 2017).

[6] To clarify, we are not arguing that child language development is best explained by innate principles of grammar that learners are believed to follow as learning progresses, as the nativist claims (Crain, 1991; Lidz et al., 2003). Our study's results do not speak directly to the validity of this approach.

REFERENCES

Abbot-Smith, K., Chang, F., Rowland, C., Ferguson, H., & Pine, J. (2017). Do two and three year old children use an incremental first-NP-as-agent bias to process active transitive and passive sentences? A permutation analysis. *PloS ONE*, *12*(10), e0186129. https://doi.org/10.1371/journal.pone.0186129

Abdelwahab, O., & Elmaghraby, A. (2016). UofL at SemEval-2016 Task 4: Multi domain word2vec for Twitter sentiment classification. In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 164–170). Association for Computational Linguistics.

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), e00938.

Adams, A. M., & Gathercole, S. E. (2000). Limitations in working memory: Implications for language development. *International Journal of Language & Communication Disorders*, *35*(1), 95–116. https://doi.org/10.1080/136828200247278

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834. https://doi.org/10.1080/03640210801929287

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

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., … 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

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The cross-linguistic study of sentence processing* (pp. 3–73). Cambridge University Press.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes*, *21*(1-3), 2–24. https://doi.org/10.1080/01690960400001721

Bever, T. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). Wiley.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, *3*(1), 19–58. https://doi.org/10.1111/j.1749-818X.2008.00099.x

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873. https://doi.org/10.1207/s15516709cog2706_2

Chan, A., Lieven, E., & Tomasello, M. (2009). Children's understanding of the agent–patient relations in the transitive construction: Cross-linguistic comparisons between Cantonese, German, and English. *Cognitive Linguistics*, *20*(2), 267–300. https://doi.org/10.1515/COGL.2009.015

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651. https://doi.org/10.1207/s15516709cog2605_3

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, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Cho, S. W. (1982). *The acquisition of word order in Korean* [Unpublished Master's thesis]. Department of Linguistics, University of Calgary.

Choi, S. (1999). Early development of verb structures and caregiver input in Korean: Two case studies. *International Journal of Bilingualism*, *3*(2/3), 241–265. https://doi.org/10.1177/13670069990030020701

Choi, Y., & Trueswell, J. C. (2010). Children's (in) ability to recover from garden paths in a verb-final language: Evidence for developing control in sentence processing. *Journal of Experimental Child Psychology*, *106*(1), 41–61. https://doi.org/10.1016/j.jecp.2010.01.003

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*(4), 597–612.

Cristante, V., & Schimke, S. (2020). The processing of passive sentences in German: Evidence from an eye-tracking study with seven-and ten-year-olds and adults. *Language, Interaction and Acquisition*, *11*(2), 163–195.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 817–828.

Church, K. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. (2019). What does BERT look at? An analysis of BERT's attention. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 276–286). Association for Computational Linguistics.

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, *67*(3), 73–97.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Dąbrowska, E. (2008). The later development of an early-emerging system: The curious case of the Polish genitive. *Linguistics*, *46*(3), 629–650. https://doi.org/10.1515/LING.2008.021

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), 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), (pp. 4171–4186). Association for Computational Linguistics.

Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants' performance on AB. *Child Development*, *56*(4), 868–883. https://doi.org/10.2307/1130099

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development*, *79*(4), 1152–1167. https://doi.org/10.1111/j.1467-8624.2008.01181.x

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2014). Familiar verbs are not always easier than novel verbs: How German pre-school children comprehend active and passive sentences. *Cognitive Science*, *38*(1), 128–151. https://doi.org/10.1111/cogs.12066

Divjak, D., Milin, P., Ez-Zizi, A., Józefowski, J., & Adam, C. (2021). What is learned from exposure: An error-driven approach to productivity in language. *Language, Cognition and Neuroscience*, *36*(1), 60–83.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Edwards, C. (2015). Growing pains for deep learning. *Communications of the ACM*, *58*(7), 14–16.

Esaulova, Y., Dolscheid, S., Reuters, S., & Penke, M. (2021). The alignment of agent-first preferences with visual event representations: Contrasting German and Arabic. *Journal of Psycholinguistic Research*, *50*(4), 843–861.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis* (pp. 1–31). Special Volume of the Philological Society. Oxford: Blackwell [Reprinted as Firth (1968)].

Futrell, R., & Levy, R. P. (2019). Do RNNs learn human-like abstract word order preferences? In G. Jarosz, M. Nelson, B. O'Connor, & J. Pater (Eds.), *Proceedings of the Society for Computation in Linguistics 2019* (pp. 50–59).

Garcia, R., & Kidd, E. (2020). The acquisition of the Tagalog symmetrical voice system: Evidence from structural priming. *Language Learning and Development*, *16*(4), 1–27. https://doi.org/10.1080/15475441.2020.1814780

Garcia, R., Rodriguez, G. G., & Kidd, E. (2021). Developmental effects in the online use of morphosyntactic cues in sentence processing: Evidence from Tagalog. *Cognition*, *216*, 104859. https://doi.org/10.1016/j.cognition.2021.104859

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691. https://doi.org/10.1111/j.1467-9280.2006.01767.x

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Erlbaum.

Givón, T. (1995). *Functionalism and grammar*. John Benjamins.

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3), 289–316. https://doi.org/10.1515/cogl.2004.011

Goldberg, Y. (2017). *Neural network methods for Natural Language Processing*. Switzerland: Springer.

Goldberg, Y., & Levy, O. (2014). Word2Vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv: Computation and Language. https://arxiv.org/abs/1402.3722

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.

Haykin, S. (2009). *Neural networks and learning machines*.Prentice Hall.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *The Proceedings of the Association for Computational Linguistics*.

Huang, Y. T., Zheng, X., Meng, X., & Snedeker, J. (2013). Children's assignment of grammatical roles in the online processing of Mandarin passive sentences. *Journal of Memory and Language*, *69*(4), 589–606. https://doi.org/10.1016/j.jml.2013.08.002

Ibbotson, P., & Tomasello, M. (2009). Prototype constructions in early language acquisition. *Language and Cognition*, *1*(1), 59–85. https://doi.org/10.1515/LANGCOG.2009.004

Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: A hierarchy of grammatical com- plexity. In F. Newmeyer & L. Preston (Eds.), *Measuring linguistic complexity* (pp. 65–82). Oxford University Press.

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. *Advances in Psychology*, *121*, 471–495.

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040.

Kemmerer, D. (2012). The Cross-Linguistic Prevalence of SOV and SVO Word Orders Reflects the Sequential and Hierarchical Representation of Action in Broca's Area. *Language and Linguistics Compass*, *6*(1), 50–66. https://doi.org/10.1002/lnc3.322

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Kim, B., Lee, Y., & Lee, J. (2007). Unsupervised semantic role labeling for Korean adverbial case. *Journal of KIISE: Software and Applications*, *34*(2), 32–39.

Kim, S. Y., Sung, J. E., & Yim, D. (2017). Sentence comprehension ability and working memory capacity as a function of syntactic structure and canonicity in 5-and 6-year-old children. *Communication Sciences & Disorders*, *22*(4), 643–656. https://doi.org/10.12963/csd.17420

Kriesel, D. (2007). *A brief introduction to neural networks*. Available at http://www.dkriesel.com

Langacker, R. W. (1991). *Foundations of cognitive grammar* (Vol. 2). Stanford University Press.

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

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151–178. https://doi.org/10.1016/S0010-0277(02)00230-5

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, *3*(4), 897–915.

Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

MacWhinney, B. (1977). Starting points. *Language*, *53*(1), 152–168. https://doi.org/10.2307/413059

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed). Lawrence Erlbaum.

Marvin, R., & Linzen, T. (2019). Targeted syntactic evaluation of language models. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 373–374.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, *3*(6), 1417–1429.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of ICLR Workshops Track*.

Mithun, N. C., Li, J., Metze, F., & Roy-Chowdhury, A. K. (2018). Learning joint embedding with multimodal cues for cross-modal video-text retrieval. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (pp. 19–27). The Association for Computing Machinery.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

O'Grady, W. (2015). Processing determinism. *Language Learning*, *65*(1), 6–32.

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.

Omaki, A., & Lidz, J. (2015). Linking parser development to acquisition of syntactic knowledge. *Language Acquisition*, *22*(2), 158–192. https://doi.org/10.1080/10489223.2014.943903

Özge, D., Küntay, A., & Snedeker, J. (2019). Why wait for the verb? Turkish speaking children use case markers for incremental language comprehension. *Cognition*, *183*, 152–180. https://doi.org/10.1016/j.cognition.2018.10.026

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

Shin, G.-H. (2021). Limits on the Agent-First strategy: Evidence from children's comprehension of a transitive construction in Korean. *Cognitive Science*, *45*(9), e13038. https://doi.org/10.1111/cogs.13038

Shin, G.-H. (2022a). Automatic analysis of caregiver input and child production: Insight into corpus-based research on child language development in Korean. *Korean Linguistics*, *18*(2), 125–158. https://doi.org/10.1075/kl.20002.shi

Shin, G.-H. (2022b). Awareness is one thing and mastery is another: Korean-speaking children's comprehension of a suffixal passive construction in Korean. *Cognitive Development*, *62*, 101184. https://doi.org/10.1016/j.cogdev.2022.101184

Shin, G.-H., & Deen, K. (2023). One is not enough: Interactive role of word order, case marking, and verbal morphology in children's comprehension of suffixal passive in Korean. *Language Learning and Development*, *19*(2), 188–212. https://doi.org/10.1080/15475441.2022.2050237

Shin, G.-H., & Mun, S. (2023). Korean-speaking children's constructional knowledge about a transitive event: Corpus analysis and Bayesian modelling. *Journal of Child Language*, *50*(2), 311–337. https://doi.org/10.1017/S030500092100088X

Sinclair, H., & Bronckart, J. P. (1972). SVO A linguistic universal? A study in developmental psycholinguistics. *Journal of Experimental Child Psychology*, *14*, 329–348. https://doi.org/10.1016/0022-0965(72)90055-0

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, *12*(3), 229–265. https://doi.org/10.1016/0010-0277(82)90033-6

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*(2), 549–565.

Sohn, H. M. (1999). *The Korean language*. Cambridge University Press.
Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child-directed speech. *Cognitive Science*, *33*(1), 75–103. https://doi.org/10.1111/j.1551-6709.2008.01004.x
Stoyneshka, I., Fodor, J. D., & Fernández, E. M. (2010). Phoneme restoration methods for investigating prosodic influences on syntactic processing. *Language and Cognitive Processes*, *25*(7-9), 1265–1293.
Stromswold, K., Pinker, S., & Kaplan, R. (1985). Cues for understanding the passive voice. *Papers and Reports on Child Language Development*, *24*, 123–130.
Sung, J., Lenz, I., & Saxena, A. (2017). Deep multimodal embedding: Manipulating novel objects with point-clouds, language and trajectories. *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2794–2801), IEEE.
Theakston, A. L., Maslen, R., Lieven, E. V. M., & Tomasello, M. (2012). The acquisition of the active transitive construction in English: A detailed case study. *Cognitive Linguistics*, *23*(1), 91–128. https://doi.org/10.1515/cog-2012-0004
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611.
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.
Vázquez, R., Raganato, A., Creutz, M., & Tiedemann, J. (2020). A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, *46*(2), 387–424.
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.
Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, *7*, 625–641.
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.
Wittek, A., & Tomasello, M. (2005). German-speaking children's productivity with syntactic constructions and case morphology: Local cues act locally. *First Language*, *25*(1), 103–125. https://doi.org/10.1177/0142723705049120
Wu, Y., Wu, W., Xing, C., Xu, C., Li, Z., & Zhou, M. (2019). A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, *45*(1), 163–197.
Xu, Y., Qiu, X., Zhou, L., & Huang, X. (2020). Improving BERT fine-tuning via self-ensemble and self-distillation. *Journal of Computer Science and Technology*, *33*(1), 1–18.
You, G., Bickel, B., Daum, M. M., & Stoll, S. (2021). Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, *11*(1), 1–11.
Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development*, *83*(4), 1382–1399. https://doi.org/10.1111/j.1467-8624.2012.01783.x

How to cite this article: Shin, G.-H., & Mun, S. (2023). Explainability of neural networks for child language: Agent-First strategy in comprehension of Korean active transitive construction. *Developmental Science*, e13405. https://doi.org/10.1111/desc.13405

APPENDIX A: Information about corpora

Name of corpus	Caregiver	Child/age range	Time of collection (year)	Quantity (sentence #)
Jiwon	M & F	Jiwon/2;0–2;3	1992	10,602
Ryu	GM, GF, & M	Jong/1;3–3;5	2009–2011	28,657
	GM, M, & F	Joo/1;9–3;10	2010–2011	27,071
	M	Yun/2;3–3;9	2009–2010	15,263

Abbreviations: F, father; GF, grandfather; GM, grandmother; M, mother.

APPENDIX B: Constructional patterns for transitive events in the caregiver input

Construction		Label	Example	Frequency	
				#	%
Canonical active transitive	No omission	Agent-first	Mia-NOM Ciwu-ACC hug	1757	25.46
	no ACC		Mia-NOM Ciwu-~~ACC~~ hug	268	3.88
	no NOM		Mia-~~NOM~~ Ciwu-ACC hug	19	0.28
Scrambled active transitive	No omission	Theme-first	Ciwu-ACC Mia-NOM hug	51	0.74
	no NOM		Ciwu-ACC Mia-~~NOM~~ hug	0	0.00
	no ACC		Ciwu-~~ACC~~ Mia-NOM hug	6	0.09
Active Transitive with omission	agent–theme, no CM	Agent-first	Mia-~~NOM~~ Ciwu-~~ACC~~ hug	3	0.04
	theme–agent, no CM	Theme-first	Ciwu-~~ACC~~ Mia-~~NOM~~ hug	0	0.00
	undetermined, no CM	Agent-first	Mia-~~NOM~~ Ciwu-~~ACC~~ hug	0	0.00
	agent–NOM only		Mia-NOM hug	935	13.55
	theme–ACC only	Theme-first	Ciwu-ACC hug	1938	28.08
	agent only, no CM	Agent-first	Mia-~~NOM~~ hug	53	0.77
	theme only, no CM	Theme-first	Ciwu-~~ACC~~ hug	1155	16.73
	undetermined, no CM[1)]	Agent-first	Mia-~~NOM~~ hug	40	0.58
Canonical suffixal passive	No omission	Theme-first	Ciwu-NOM Mia-DAT hug-psv	2	0.03
	no DAT		Ciwu-NOM Mia-~~DAT~~ hug-psv	0	0.00
	no NOM		Ciwu-~~NOM~~ Mia-DAT hug-psv	0	0.00
Scrambled suffixal passive	No omission	Agent-first	Mia-DAT Ciwu-NOM hug-psv	1	0.01
	no NOM		Mia-DAT Ciwu-~~NOM~~ hug-psv	0	0.00
	no DAT		Mia-~~DAT~~ Ciwu-NOM hug-psv	0	0.00
Suffixal passive with omission	theme–agent, no CM	Theme-first	Ciwu-~~NOM~~ Mia-~~DAT~~ hug-psv	0	0.00
	agent–theme, no CM	Agent-first	Mia-~~DAT~~ Ciwu-~~NOM~~ hug-psv	0	0.00
	undetermined, no CM	Theme-first	Ciwu-~~NOM~~ Mia-~~ACC~~ hug-psv	0	0.00
	theme–NOM only		Ciwu-NOM hug-psv	407	5.90
	agent–DAT only	Agent-first	Mia-DAT hug-psv	13	0.19
	theme only, no CM	Theme-first	Ciwu-~~NOM~~ hug-psv	20	0.29
	agent only, no CM	Agent-first	Mia-~~DAT~~ hug-psv	0	0.00
	undetermined, no CM[2)]	Theme-first	Mia-~~NOM~~ hug-psv	0	0.00
Ditransitive	recipient–DAT only[1)]	Agent-first	Mia-DAT give	234	3.39
		SUM		6902	100.00

Note: CM = case-marking. *Ciwu* and *Mia* are human names. 1) and 2) were determined by the typical thematic role ordering in each construction type expressing transitive events: agent-before-theme for 1) (active transitive); theme-before-agent for 2) (suffixal passive). We included a ditransitive construction with only a recipient–dative pairing. Although it does not relate to a transitive event per se and does not count as a relevant pattern, we considered this constructional pattern here because the dative marker is often used to indicate a recipient in the active and thus a potential competitor of the agent–dative pairing in the passive.