Limits on Neural Networks: Agent-First Strategy in Child Comprehension

Gyu-Ho Shin (gyuho.shin@upol.cz)
Department of Asian Studies, Palacký University Olomouc
tř. Svobody 26, Olomouc, Czech Republic

Seongmin Mun (email address)
Department of English Language and Literature, Chosun University
309 Pilmun-daero, Seoseok-dong, Dong-gu, Gwangju, Republic of Korea

Abstract

This study investigates how neural networks reveal developmental trajectories of child language, focusing on the *Agent-Firs*t strategy in comprehension of an active transitive construction in Korean. We develop three models (LSTM; BERT; GPT-2) and measure their classification performance on the test stimuli used in Shin (2021) involving scrambling and omission of constructional 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: Agent-First strategy; Neural network; Active transitive; Child comprehension

Introduction

One active trend in research on child language development is to adopt computational modelling techniques to address developmental trajectories of children's linguistic knowledge (e.g., Alishahi & Stevenson, 2008; Ambridge et al., 2020; Bannard et al., 2009; You et al., 2021). There is growing interest in the ways neural networks (NNs) address human language behaviour (e.g., Futrell & Levy, 2019; 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 Kriesel, 2007 for in-depth descriptions of neural networks). NNs are applied to various disciplines (Abiodun et al., 2018) due to its efficient performance on data analysis factors (Wang et al., 2017), but they require exceedingly large training samples and considerable computing resources for effective operation (e.g., Edwards, 2015). Moreover, the continuous development of NN algorithms has made their internal mechanisms deviate from how biological neurons operate in reality (e.g., Crick, 1989). Recent studies have shown that transformers, which are characterised as the attention mechanism (e.g., Vaswani et al., 2017), yield better performance on language tasks than previously proposed architectures (e.g., Hawkins et al., 2020). Nevertheless, we are not aware of any study that attempts to explain properties of child language through the lens of NNs, particularly any that reveals the extent to which NN models address the findings of behavioural experiments around children.

The present study investigates this inquiry, focusing on the *Agent-First* strategy in child comprehension. Children often map the first noun (mostly the subject) of a sentence to an agent role during comprehension (e.g., Abbot-Smith et al., 2017; Sinclair & Bronckart, 1972; Slobin & Bever, 1982). 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 (e.g., 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). Existing literature, mostly based on the major languages currently under investigation, reports children's heavy reliance on this strategy for sentence comprehension (e.g., Abbot-Smith et al., 2017; Gertner et al., 2006; 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.

We pursue this inquiry through an active transitive construction in Korean, an agglutinative, SOV language with overt case-marking. The canonical word order for the active transitive follows agent-theme ordering as in (1a); this can be scrambled as in (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). We develop various

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

NN models—*LSTM* (Hochreiter & Schmidhuber, 1997), *BERT* (Devlin et al., 2018), and *GPT-2* (Radford et al., 2019)—to explore what these models can(not) demonstrate about child comprehension with respect to the *Agent-First* strategy. We train each model with caregiver input in the CHILDES database (MacWhinney, 2000), and evaluate its performance in classifying test items used in Shin (2021).

(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.'

Korean-speaking children's utilisation of the *Agent-First* strategy to comprehension of active transitive

Shin (2021) 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 three-to-six-year-old children's comprehension of the active transitive construction involving scrambling and omission of constructional components through a series of picture-selection tasks. To this end, Shin devised an innovative methodology that systematically obscured parts of test stimuli with acoustic masking (e.g., coughing, chewing, yawning) accompanied by child-friendly contexts.

Shin (2021) notes four major findings (Table 2). First, whereas the children had a good command of case-marking knowledge (NOM indicating the agent; ACC indicating the theme), they showed asymmetric performance by canonicity: they were better in the canonical ($N_{NOM}N_{ACC}V$) than the scrambled ($N_{ACC}N_{NOM}V$) condition. Second, they did not manifest the agent-first interpretation strongly in $N_{\cancel{CASE}}V$, showing around 40 per cent for the 3-4yrs and around 60 per cent for the 5-6yrs. In this condition, children must determine the thematic role of the first and 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 $N_{\cancel{CASE}}V$, the presence of a second noun ($N_{\cancel{CASE}}N_{\cancel{CASE}}V$) increased responses consistent with the *Agent-First* strategy, but its magnitude differed by age: only the 3-4yrs considerably enhanced the agent-first interpretation from $N_{\cancel{CASE}}V$ to $N_{\cancel{CASE}}N_{\cancel{CASE}}V$. Fourth, the presence of markers ($N_{NOM}V$) substantially increased the agent-first response rates for both age groups.

Based on these findings, Shin (2021) 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 participants (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 agent–NOM association). The other force involves the developing nature of a child processor, prioritising a local cue over a distributional cue (Wittek & Tomasello, 2005). Children may attend to the local pairing that associates the NOM-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 complete interpretation of a transitive sentence. Because the activation of the *Agent-First* strategy is tied to other grammatical cues such as case-marking (particularly NOM) and a second nominal, 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 comprehension bias, as claimed by previous studies targeting the major languages (e.g., Abbot-Smith et al., 2017; Gertner et al., 2006).

Table 1. Summary of results: major conditions (α = .05)

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

[1] Abbreviation: ACC = accusative case marker; CASE = case marker (unspecified); NOM = nominative case marker; PST = past tense marker; SE = sentence ender; V = verb.

With these in mind, we ask whether/how NNs, as a proxy for children's cognitive space wherein learning occurs, reveal their developmental trajectories as a function of the interplay between properties of input (child-directed speech) and domain-general learning capacities (statistical learning). We pursue this inquiry by developing three NN models with caregiver input and measuring their classification performance on the same stimuli used in Shin (2021), specifically focusing on the major conditions relating to the *Agent-First* strategy listed in Table 1. For model training, we employ the caregiver input data extracted from CHILDES pertaining to transitive events to reflect the experimental setting of Shin (2021), where children's interpretation was contextualised through pictures before presenting aural stimuli. It is known that 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) (e.g., Cameron-Faulkner et al., 2003)—effectively supports children's development of linguistic knowledge (e.g., Behrens, 2006; Choi, 1999). If NNs faithfully respect this characteristic, the models in this study should approximate child comprehension patterns measured by Shin (2021), with reasonable accuracy, like their successful performance in some adult language features (e.g., Futrell & Levy, 2019; Hawkins et al., 2020; Warstadt & Bowman, 2020).

Methods

With the *Python* packages and pre-trained models, we trained three models (Table 2) with all the caregiver input data in CHLDES, along with parameter setting advised by previous studies (e.g., Vázquez et al., 2020; Wu et al., 2019). The caregiver input data 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. These treatments resulted in 69,498 sentences (285,350 words).

Table 2. Summary: Model specification

	LSTM	BERT	GPT-2
Package	*PyTorch*	*Transformers*	
Pre-trained model	*KoCharElectra-Base* (Park, 2020); 11,568 syllable types	*KoBERT* (Jeon, Lee et al., 2019); 54-million-word tokens	*KoGPT2-base-v2* (Jeon, Kim et al., 2019); 51,200-word tokens
Tokeni-sation	Syllable-based	Syllable-based *WordPiece*	Syllable-based *Byte Pair Encoding*
Model-specific	Hidden layers: 256, Epoch: 10, Learning rate: .00002	Batch: 32, Sequence length: 256, Epsilon: .00000001, Seed: 42, Epoch: 30, Learning rate: .0001	

We note that we used the respective pre-trained models in developing each NN model. While NNs typically require large training data for their optimal operation, 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 use that children experience. Adopting a pre-trained model in conjunction with the caregiver input data can be one way to approximate this nature, possibly ensuring better ecological validity for the simulation. Notably, no research has ever touched upon this issue, thus worthy of further attention.

Table 3. Constructional patterns for transitive events in the caregiver input (adapted from Shin, 2020)

Construction		Label	Frequency #	Frequency %
Canonical active transitive	No omission	Agt-1st	1,757	25.46
	no ACC		268	3.88
	no NOM		19	0.28
Scrambled active transitive	No omission	Thm-1st	51	0.74
	no NOM		0	0.00
	no ACC		6	0.09
Active Transitive with omission	agent–theme, no CM	Agt-1st	3	0.04
	theme–agent, no CM	Thm-1st	0	0.00
	undetermined, no CM	Agt-1st	0	0.00
	agent–NOM only		935	13.55
	theme–ACC only	Thm-1st	1,938	28.08
	agent only, no CM	Agt-1st	53	0.77
	theme only, no CM	Thm-1st	1,155	16.73
	undetermined, no CM[1)]	Agt-1st	40	0.58
Canonical suffixal passive	No omission	Thm-1st	2	0.03
	no DAT		0	0.00
	no NOM		0	0.00
Scrambled suffixal passive	No omission	Agt-1st	1	0.01
	no NOM		0	0.00
	no DAT		0	0.00
Suffixal passive with omission	theme–agent, no CM	Thm-1st	0	0.00
	agent–theme, no CM	Agt-1st	0	0.00
	undetermined, no CM	Thm-1st	0	0.00
	theme–NOM only		407	5.90
	agent–DAT only	Agt-1st	13	0.19
	theme only, no CM	Thm-1st	20	0.29
	agent only, no CM	Agt-1st	0	0.00
	undetermined, no CM[2)]	Thm-1st	0	0.00
Ditransitive	recipient–DAT only[1)]	Agt-1st	234	3.39
Sum			6,902	100.00

Note. CM = case-marking. *Ciwu* and *Mia* are human names. The labels of 1) and 2) were determined by the typical thematic role ordering in each construction type. 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.

For the binary classification of test items (Agent-First; Theme-First) in consideration of the experimental setting of Shin (2021), 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 (Table 3). The instances were extracted from the pre-processed caregiver input data through an automatic search process (cf. Shin, 2020); every sentence for each extraction was 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, 2020). Furthermore, considering the zero occurrence of some patterns in the input, we adapted the Laplace smoothing technique (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 per cent of the entire input.

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 4). Each trained model classified every test stimulus, evaluating if the stimulus fell into Agent-First or Theme-First. We note that, while the stimuli of $N_{\cancel{CASE}}N_{\cancel{CASE}}V$ and $N_{\cancel{CASE}}V$ in Shin (2021) involved acoustic masking, the same stimuli types 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.

Table 4. Composition of test stimuli

Condition	Example	Expected classification
$N_{NOM}N_{ACC}V$	dog-NOM cat-ACC poke	Agent-first
†$N_{ACC}N_{NOM}V$	cat-ACC dog-NOM poke	Theme-first
$N_{NOM}V$	dog-NOM poke	Agent-first
$N_{ACC}V$	cat-ACC poke	Theme-first
$N_{\cancel{CASE}}N_{\cancel{CASE}}V$	dog cat poke	Agent-first
$N_{\cancel{CASE}}V$	dog poke	Agent-first

There is no syllable-based Korean pre-trained model exclusively for LSTM, so we adapted a pre-trained model for ELECTRA to extract relevant vocabulary information to train our LSTM model. We separated sentences from labels in the caregiver input data and tokenised the sentences by syllable, imitating the structure of the pre-trained model. All the syllables obtained from the pre-trained model and the caregiver input data were submitted to the model's input layer, and the number of sentence labels were utilised as its output layer. Once the training was completed, the model processed the test stimuli, accumulating by-syllable information sequentially (by generating respective hidden layers), and it compared the outcomes (as a value of one for Agent-First or zero for Theme-First) to the actual labels of these stimuli. We repeated the same learning process 30 times and averaged the by-condition outcomes to assess the models' classification performance, controlling for potential unexpected variations from any training phase.

For the BERT model, 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. A 'Label' column was added to indicate whether the sentence was Agent-first or Theme-first. We tokenised the sentences by syllable (mirroring the pre-trained model) and converted them into numeric values which served as designated indices of the tokens in the pre-trained model. All the information obtained by this process was transformed into a *tensor* (i.e., a data format reducing the size to make processing faster). The initial values of epsilon, learning rate, and seed were automatically updated with the outcomes of each epoch. The training occurred 960 times (32 batches * 30 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). The trained model classified the test stimuli; like 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 the GPT-2 model used no symbol to mark the start/end of each input sentence. 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. Notably, however, both *KoBERT* and *KoGPT-2* employed 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.

Results and Discussions

Case-marked conditions

Figure 1 illustrates the classification performance of the three models, together with the children's and adults' performance measured in Shin (2021), on the four case-marked conditions. For the two-argument conditions, each model demonstrated asymmetric rates of accuracy. The LSTM model was constantly at-ceiling for both conditions (M = 90.28, SD = 0.30 for $N_{NOM}N_{ACC}V$; M = 91.67, SD = 0.28 for $N_{ACC}N_{NOM}V$), approximating the adults' accuracy rates. In contrast, the other two models' performance was affected by canonicity: they showed a drop in accuracy for the scrambled condition relative to the canonical counterpart (BERT: M = 100.00, SD = 0.00 for $N_{NOM}N_{ACC}V$; M = 51.61, SD = 0.50 for $N_{ACC}N_{NOM}V$; GPT-2: M = 100.00, SD = 0.00 for $N_{NOM}N_{ACC}V$; M = 16.67, SD = 0.37 for $N_{ACC}N_{NOM}V$). This trend was somewhat similar to the children's performance, but the gap between the two conditions was much larger for the models than for the children. For the one-argument conditions, all the models achieved above-chance performance (LSTM: $M = 72.22$, $SD = 0.45$ for $N_{NOM}V$; $M = 100.00$, $SD = 0.00$ for $N_{ACC}V$; BERT: $M = 85.00$, $SD = 0.36$ for $N_{NOM}V$; $M = 97.22$, $SD = 0.16$ for $N_{ACC}V$; GPT-2: $M = 83.33$, $SD = 0.37$ for $N_{NOM}V$; $M = 83.33$, $SD = 0.37$ for $N_{ACC}V$), which resembled the children's accuracy rates.

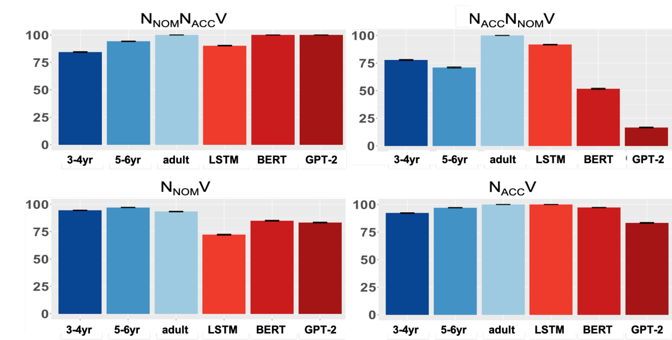

Figure 1. Child comprehension and model performance. X-axis: group (for experiment) or model (for simulation). Y-axis: accuracy. Error bars indicate 95% CI.

These results suggest two major aspects in the models' classification performance on the case-marked conditions. First, it seems that BERT and GPT-2 followed characteristics of the caregiver input selectively. Recall that (i) the number of first-noun-as-agent patterns (3,049 instances) did not exceed that of first-noun-as-theme patterns (3,579 instances) and (ii) the number of nominative-first patterns (overtly marked with the nominative case marker; 3,369 instances) outnumbered that of accusative-first patterns (overtly marked with the accusative case marker; 1,989 instances) despite the generally higher omission rate of the accusative case marker than that of the nominative case marker in caregiver input (Shin, 2020). Given these properties, the three 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. This may have led 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 model performance 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).

Moreover, the LSTM model's outperformance over the other two transformer-architecture models in $N_{ACC}N_{NOM}V$—against our prediction—indicates the algorithm-exclusive memory cell's contribution to information processing. In other words, the existence of a memory cell may have assisted the classification accuracy as effectively as the attention mechanism in the transformer-architecture models in the given simulation environment. Considering that transformer architecture excels in utilising information from long input sequences, it is reasonable to think that BERT and GPT-2 may not have fully exerted their algorithmic strength when handling child language. The LSTM model's good classification performance further aligns with previous reports on this model's success in learning and generalising clause-level linguistic knowledge (Futrell & Levy, 2019; Wilcox et al., 2018). In particular, 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.

Case-less conditions

Figure 2 illustrates the classification performance of the three models, together with the children's and adults' performance measured in Shin (2021), on the two case-less conditions. 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. For $N_{CASE}N_{CASE}V$, the LSTM model was above-chance ($M = 63.89$, $SD = 0.48$), and the BERT and GPT-2 models were below-chance (BERT: $M = 34.44$, $SD = 0.48$; GPT-2: $M = 33.33$, $SD = 0.47$). For $N_{CASE}V$, all the models were below-chance (LSTM: $M = 25.00$, $SD = 0.43$; BERT: $M = 18.89$, $SD = 0.39$; GPT-2: $M = 0.00$, $SD = 0.00$).

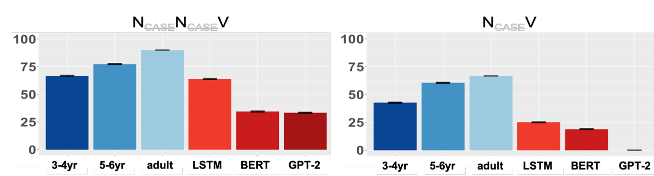

Figure 2. Child comprehension and model performance. X-axis: group (for experiment) or model (for simulation). Y-axis: proportion of agent-first interpretation. Error bars indicate 95% CI.

These results indicate that the models failed to capture the trend manifested by the children. The BERT and GPT-2 models malfunctioned in these conditions, performing with high deviation from the children's interpretation for the same conditions. The performance of the LSTM model was close to the children's interpretation in $N_{CASE}N_{CASE}V$, but differed considerably from it in $N_{CASE}V$. 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. The under-informativeness in determining the thematic role of the first noun involving the two conditions would have affected both the children's comprehension and these models' performance. However, the three models may have been more influenced than the children by the lack of reference point for the classification decision (i.e., case-marking) in the stimuli, rendering their performance substantially deviant from the children's response rates. Notably, compared to $N_{CASE}V$, the LSTM model improved its performance towards Agent-First when additional information (a second nominal) appeared in $N_{\cancel{CASE}}N_{\cancel{CASE}}V$. This improvement is ascribable to the same reason suggested for its performance on the case-marked conditions: a memory cell may have helped this model better utilise this additional information than the attention mechanism in their search for the intended label of this condition.

General Discussion

This study's results can be attributed to various factors. For instance, the simulation environment in this study may not have perfectly conformed to the experimental setting of Shin (2021) to the extent that the models utilised relevant information from the stimuli in the exact same way as the children did in the experiment. We trained each model with all the transitive-event instances in CHILDES, reflecting how the children in Shin (2021) attuned their interpretation to transitive events before they were exposed to the stimuli. Despite this treatment, our models might not have had a testing environment fully compatible with the one that the children experienced. Moreover, while the experimental stimuli in Shin (2021) employed acoustic masking to obscure the case markers so the children would notice that there was something but hidden, the same stimuli in the simulation involved no such acoustic signals. 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 have affected the model performance in an unexpected way. 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 yet in this field, we cannot say for certain that these are the all-and-only reason of this asymmetry.

Another possible factor for these models' odd performance is around language-specific properties. While Korean caregiver input joins the general characteristics of child-directed speech (e.g., Cameron-Faulkner et al., 2003), it also manifests language-specific properties such as scrambling and omission of sentential components at varying levels. In addition to the general nature of caregiver input, the NN models may have thus been affected by the specific word order and/or the presence of case markers in conducting the classification, as shown with the two-argument case-marked scrambled condition ($N_{ACC}N_{NOM}V$) and the two case-less conditions ($N_{\cancel{CASE}}N_{\cancel{CASE}}V$; $N_{\cancel{CASE}}V$). This aligns with previous reports on language-specific challenges for automatic processing of Korean (e.g., Shin & Jung, 2021; Kim et al., 2007). Since we are unaware of any study on the contribution of language-specific properties to NNs' performance on child language, this claim awaits further examination.

In addition to these factors, we argue that the characteristics of these models' internal algorithms 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 a particular formal sequence involving words/characters; that is, they rely heavily on form. This yields a context in a computational sense, but importantly, it is qualitatively different from a linguistic context, which involves semantic–pragmatic considerations. Hence, whenever the models access the meaning/function of a linguistic unit, they exploit the formal co-occurrence in the incoming input, rather than directly drawing upon the meaning/function of the linguistic unit of interest during their processing. Moreover, NNs are designed to generalise what they already have (through pre-trained models and information from the training), but are not designed to make reasonable predictions outside of a trained range (Ye, 2020). This algorithmic nature—which exclusively utilises sequence-based formal information existent within a model—may have rendered the models in this study deviant from the children's performance on some test stimuli possibly out of range. The key evidence comes from the models' performance on $N_{\cancel{CASE}}V$ (the condition in which 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 $N_{ACC}V$ (the conditions in which the same learner has more, and core, information about the first noun's thematic role indicated by specific case markers next to the noun).

This manner of algorithmic operation differs from how a human processor deals with linguistic knowledge, which is characterised as simultaneous activation of multiple (non-)linguistic routes in parallel and immediate mapping of form onto function (and vice versa) to reduce the burden of work at hand (e.g., Karimi & Ferreira, 2016; McRae & Matsuki, 2009; O'Grady, 2015; Traxler, 2014), despite the same pursuit of efficiency in information processing like a computation model. In particular, considering the developing nature of a child processor (e.g., Choi & Trueswell, 2010; Omaki & Lidz, 2015; Özge et al., 2019; Snedeker & Trueswell, 2004), the children in Shin (2021) may have made the best (albeit imperfect) use of the information available at the time, based on their learning trajectories. That is, when they computed the relative agenthood between the two arguments with no animacy hierarchy involved, their interpretation may have been swayed away by multiple sources, including verb semantics, event/world knowledge, and cognitive bias such as the *Agent-First* strategy.

To conclude, while NNs tested in this study (and perhaps any currently developed computational algorithms) can utilise information about formal co-occurrences to access the intended message to a certain degree, the outcome of this process may be substantially different from how a child (as a developing processor) engages in comprehension. Despite this study's simulation-wise limitations, the implications of the current study provide evidence of some limitation in the NNs' capacity for revealing developmental trajectories of child language.

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.

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

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834.

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

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*(41), 17284–17289.

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). New York: Cambridge University Press.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and Cognitive Processes*, *21*(1-3), 2–24.

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.

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

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.

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.

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.

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).

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.

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

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*. London: 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.

Jeon, H., Kim, H., Jung, S., Kim, M., Maeng, Y., Kang, K., & Moon, S. (2019). *KoGPT2 Ver 2.0.* Retrieved from https://github.com/SKT-AI/KoGPT2 on 15-September-2021

Jeon, H., Lee, D., & Park, J. (2019). *Korean BERT pre-trained cased (KoBERT)*. Retrieved from https://github.com/SKTBrain/KoBERT on 15-September-2021

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.

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.

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

MacWhinney, B. (1977). Starting points. *Language, 53*(1), 152–168.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd edition). Mahwah, NJ: Lawrence Erlbaum.

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.

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

Omaki, A., & Lidz, J. (2015). Linking parser development to acquisition of syntactic knowledge. *Language Acquisition*, *22*(2), 158–192.

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.

Park, J. (2020). *Korean syllable-based vocabulary*. Retrieved from https://github.com/monologg/KoCharELECTRA/blob/master/vocab.txt on 12-October-2021.

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. (2020). *Connecting input to comprehension: First language acquisition of active transitives and suffixal passives by Korean-speaking preschool children* [Unpublished Ph.D. dissertation]. Department of Linguistics, University of Hawaiʻi at Mānoa.

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.

Shin G-H., & Jung, B. K. (2021a). Automatic analysis of learner corpora in Korean: Written production of Korean passive constructions for Mandarin-speaking learners of Korean. *International Journal of Learner Corpus Research, 7*(1), 53–82.

Sinclair, H., & Bronckart, J. P. (1972). SVO A linguistic universal? A study in developmental psycholinguistics. *Journal of Experimental Child Psychology*, *14*, 329–348.

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.

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.

Sohn, H. M. (1999). *The Korean language*. New York, NY: Cambridge University Press.

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.

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.

Ye, A. (2020). You don't understand neural networks until you understand the Universal Approximation Theorem: The proof behind the neural network's power. Retrieved from https://medium.com/analytics-vidhya/you-dont-understand-neural-networks-until-you-understand-the-universal-approximation-theorem-85b3e7677126 on 05-DEC-2021.

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.