From: **Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, and Veronica Ramenzoni** <cogsci22a@precisionconference.com>
Date: Tue, 12 Apr 2022 at 17:26
Subject: Decision on submission 1358 to CogSci 2022
To: Gyu-Ho Shin <gyuhoshin@gmail.com>

Dear Gyu-Ho Shin :

We are very pleased to inform you that your paper submission

1358 - Limits on Neural Networks: Agent-First Strategy in Child Comprehension

has been accepted for presentation as a poster with full paper publication at CogSci 2022. We received 762 paper submissions this year, and each underwent careful peer review. We accepted 205 (27%) for oral presentation, 337 (44%) for poster presentation with full paper publication and 151 for poster presentation with abstract publication (20%). Please read the remainder of this entire email carefully, as it contains important information about your presentation, and instructions for finalizing your submission for publication in the proceedings.

We are very pleased to offer a fully hybrid CogSci 2022. This year, you have the choice of presenting a traditional poster in person, or giving an online flash talk (an alternative to traditional posters, consisting of a 4-5 minute virtual talk in a themed session, with some time for discussion).

Full guidelines and instructions for presentation, both in person and virtual, will be made available to presenters around mid-May. When you submit your final paper or abstract, we will ask you to tell us whether you plan to present the poster virtually or in person. If you plan to present virtually, then you will be allocated to a “flash talk” session [see below], and so you will need to indicate what timezone the person presenting it will be in for the conference. If you plan to present in-person, you can choose between a physical poster or participating in a flash talk session; we will ask you to indicate this as well. We need this information so that we can create an appropriate schedule; if you change your mind we will do our best to accommodate it but cannot guarantee that the session you are placed in is appropriate to your topic.

Details concerning the timing of your poster in the conference program will be provided after May 11, 2022 in the coming weeks.

To make the hybrid conference work successfully, we will have premium A/V setup and personnel to ensure that the streaming is of high quality and the interactions between the in-person and online participants are smooth. All session blocks will contain a mix of talks, traditional posters, and flash talks. This ensures that people in all time zones will have at least a few sessions where they can stream in live. All talk sessions will be recorded and pdfs of posters will be made accessible online to all participants.

The dates of the conference are July 27 – 30, 2022. Detailed conference schedules and guidelines about participation in the conference will be made available through the conference webpage:

https://cognitivesciencesociety.org/cogsci-2022/

You have the option of publishing either your full paper or just the abstract in the CogSci Conference Proceedings. This option was introduced to address concerns that some authors have expressed about publishing the same, or a similar, paper in a journal after it has already appeared in the CogSci Proceedings. Although the Cognitive Science Society's policy is clear that proceedings publications should not block journal publications, there have been isolated issues with some journals.

If you decide to publish the full paper, please take into account in your revisions the reviewers' comments that appear at the end of this message. Your final paper can be no longer than six pages, plus unlimited space for acknowledgements and references, formatted using one of the templates found here:

https://dearmond.sharepoint.com/:f:/g/ExternalSharing/EgcbkEToDqpDojTHPw0XgTMBPueRKMZADepPevXiBA8LCA?e=7dKKRb

It is critical that you use one of the templates for the final version without altering font size, spacing, margins, etc. Any papers with format alterations, or exceeding 6 pages (not including acknowledgments and references) will be ** rejected ** and will not be published, nor will a presentation be scheduled.

Regardless of whether you choose to publish the full paper or just the abstract, you must login to the Precision Conference website (link below) to make your final submission. If you do not take this final step, neither your paper nor your abstract will appear in the conference proceedings. Thus, it is very important that you complete this step before May 11, 2022 at:

https://new.precisionconference.com/cogsci

Details about conference registration and registration fees are available here:

https://cognitivesciencesociety.org/registration/

Registration fees will be offered at substantially reduced rates for Cognitive Science Society members. In an effort to be more inclusive and global, CSS has a “Pay what you can (PWYC)” membership rate. More information regarding membership can be found on the website:

https://cognitivesciencesociety.org/membership/

In order for your poster to be presented and to appear in the proceedings, at least one of the authors must be registered for the conference.

Thank you very much for contributing your interesting work to CogSci 2022. We look forward to seeing you at the hybrid conference in July!

With best regards,
Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, and Veronica Ramenzoni
cogsci@podiumconferences.com

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

Anonymous CogSci submission

Abstract

This study investigates how neural networks address child language development, specifically focusing on the *Agent-Firs*t strategy in comprehension of an active transitive construction in Korean. We develop various models (LSTM; BERT; GPT-2) and measure their classification performance on the test stimuli used in Author (xxxx) 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

For comprehension, children often map the first noun (mostly the subject) of a sentence to an agent role (e.g., Abbot-Smith et al., 2017; Sinclair & Bronckart, 1972; Slobin & Bever, 1982). This comprehension bias, dubbed the *Agent-First* strategy, 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). It has also been observed that the first item in a sequence holds a privileged status in human cognition; language users employ the first element in a sentence as a starting point for language behaviour, guiding the rest of the sentence (e.g., 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 (e.g., 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 (e.g., Bornkessel-Schlesewsky & Schlesewsky, 2009; Cohn & Paczynski, 2013). Existing literature reports children's heavy reliance on this strategy for sentence comprehension (e.g., Abbot-Smith et al., 2017; Gertner et al., 2006; Huang et al., 2013). This favours the early emergence and universal application of this strategy as an intrinsic cognitive bias for child comprehension across languages.

A recent trend in child language research is the adoption of computational methods to address developmental trajectories of linguistic knowledge (e.g., Alishahi & Stevenson, 2008; Ambridge et al., 2020; Bannard et al., 2009; You et al., 2021). There is a growing interest in the ways neural network (NN) models address human language behaviour (e.g., Futrell & Levy, 2019; Warstadt & Bowman, 2020; Warstadt et al., 2019), under the assumption that artificial NNs reflect how biological NNs operate in human brains (Elman, 1990; Haykin, 2009; Hopfield, 1982). Recent studies have shown that transformer architecture yields better performance on language tasks than previously proposed architectures (e.g., Hawkins et al., 2020; Vaswani et al., 2017). Nevertheless, we are not aware of any study that investigates child language development through the lens of NNs (as a proxy for learner's cognitive space for learning), particularly any that reveals the extent to which NN models account for the findings of behavioural experiments around children.

Against this background, the present study investigates how NNs address child language, with a particular focus on the *Agent-First* strategy in comprehension of an active transitive construction in Korean, an understudied language for this topic. We develop various NN models—Long Short-Term Memory (*LSTM*; Hochreiter & Schmidhuber, 1997), Bidirectional Encoder Representations from Transformers (*BERT*; Devlin et al., 2018), and Generative Pre-trained Transformer 2[1] (*GPT-2*; Radford et al., 2019)—to explore what these models can(not) demonstrate about child comprehension regarding this strategy.

Korean-speaking children's comprehension of active transitive construction

Korean is an agglutinative, Subject–Object–Verb language with overt case-marking through dedicated markers. The canonical word order for the active transitive follows agent–theme ordering (1a); this can be scrambled (1b), manifesting

[1] This model was further improved and proposed as GPT-3 (Floridi & Chiriatti, 2020), but its source code is not open-access at the moment of study.

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 or a combination of an argument and a case marker 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[2]
'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.'

Author (xxxx), the reference research for the current study, 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 such as case-marking and the presence of a second nominal and does not function independently of them. Author 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, Author devised an innovative methodology that systematically obscured parts of test stimuli with acoustic masking (e.g., coughing, chewing, yawning) accompanied by child-friendly contexts.

Author notes four major findings (Table 2). 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 asymmetric performance by canonicity: they were better in the canonical condition ($N_{NOM}N_{ACC}V$) than the scrambled condition ($^{\dagger}N_{ACC}N_{NOM}V$). Second, they did not manifest the agent-first interpretation strongly in $N_{CASE}V$, showing at-chance performance for the 3-4yrs and weakly above-chance performance 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 $N_{CASE}V$, 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: only the 3-4yrs considerably enhanced the agent-first interpretation from $N_{CASE}V$ to $N_{CASE}N_{CASE}V$. Fourth, the presence of markers substantially increased the agent-first response rates for both age groups, as shown in $N_{NOM}V$.

Table 1. Summary of results: major conditions ($\alpha = .05$)

Condition	Group	Mean (%)	SD	Chance level
$N_{NOM}N_{ACC}V^{(a)}$	3-4yr	84.44	0.36	Above-chance***
	5-6yr	94.20	0.24	Above-chance***
	Adult	100.00	0.00	Above-chance***
$^{\dagger}N_{ACC}N_{NOM}V^{(a)}$	3-4yr	77.78	0.42	Above-chance***
	5-6yr	71.01	0.46	Above-chance***
	Adult	100.00	0.00	Above-chance***
$N_{NOM}V^{(a)}$	3-4yr	94.44	0.23	Above-chance***
	5-6yr	97.10	0.17	Above-chance***
	Adult	93.33	0.25	Above-chance***
$N_{ACC}V^{(a)}$	3-4yr	92.22	0.27	Above-chance***
	5-6yr	97.10	0.17	Above-chance***
	Adult	100.00	0.00	Above-chance***
$N_{CASE}N_{CASE}V^{(b)}$	3-4yr	66.67	0.48	Above-chance**
	5-6yr	77.27	0.42	Above-chance***
	Adult	90.00	0.04	Above-chance***
$N_{CASE}V^{(b)}$	3-4yr	42.59	0.50	At-chance
	5-6yr	60.42	0.49	Above-chance*
	Adult	66.67	0.06	Above-chance**

Note. While the scoring for (a) indicates accuracy (1: correct; 0: incorrect), that for (b) indicates the high likelihood of agent-first interpretation (1: agent-first; 0: theme-first) as (b) can in principle be interpreted in more than one way.

Based on these findings, Author concludes 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. 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), so Korean-speaking children employ this strategy with confidence only when they are provided with a linguistically informative environment. This argument challenges the long-standing claim that children have the default mapping of the agent onto the first noun as an intrinsic bias for comprehension (e.g., Abbot-Smith et al., 2017; Gertner et al., 2006; Huang et al., 2013).

Author argues that, given the 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 the properties of caregiver input regarding transitive events. In CHILDES (MacWhinney, 2000), 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,

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

prioritising a local cue over a distributional cue (Wittek & Tomasello, 2005) given that the processor must deal with various (non-)grammatical cues during comprehension. 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 *Agent-First* strategy for a complete interpretation of the sentence at hand.

With these in mind, the present study asks whether and 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 measure three NN models' classification performance on the same stimuli used in Author (xxxx), paying particular attention to 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 Author (xxxx), where children's interpretation was contextualised through pictures before presenting aural stimuli. It is known that caregiver input—which is simple, short, and repetitive by nature in comparison to adult speech (e.g., Cameron-Faulkner et al., 2003)—effectively supports children's development of linguistic knowledge (e.g., Behrens, 2006; Snow, 1972). If NNs faithfully respect this characteristic, the models should approximate the children's comprehension behaviour measured by Author (xxxx), with reasonable accuracy, like their successful performance in adult language (e.g., Futrell & Levy, 2019; Hawkins et al., 2020; Warstadt & Bowman, 2020). Due to the transformer architecture's better capability of capturing human language behaviour than the recurrent architecture (e.g., Hawkins et al., 2020), BERT and GPT-2 should be closer than LSTM in classification performance relative to children's performance in Author (xxxx).

Methods

Composition of input and test items

We included in the input all the constructional patterns expressing transitive events—active transitive and suffixal passive, with scrambling and varying degrees of omission manifested—found in CHLDES. The raw data proceeded first to a pre-processing stage: typos and spacing errors were corrected; any sentence whose length was less than five characters or those consisting only of onomatopoeia and mimetic words were excluded; any non-verb-final instance (e.g., *Yengswu*-NOM read-SE book-ACC; eat-SE rice-ACC) was also excluded from the data (see Author, xxxx for the details about the pre-processing). These treatments resulted in 69,498 sentences (285,350 words) for the caregiver input. The pre-processed data were then inputted to an automatic search process to extract instances of the two construction types. Every list of sentences for each extraction was also checked manually to ensure its accuracy. Table 2 illustrates the final caregiver input composition for model training.

Table 2. Input composition for model training: constructional patterns for transitive events in the caregiver input in CHILDES (adapted from Author, xxxx)

Construction		Label	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.

Although this study's main focus was the active transitive, we included the suffixal passive in the input. This construction is another major clause-level device expressing a transitive event and is known to be the most frequent type of passive in caregiver input (Author, xxxx). The suffixal passive thus serves as the representative type of passive that children are likely to encounter regarding a transitive event. We also included a ditransitive construction with only a recipient–dative pairing. Although it does not relate to a transitive event, we considered it 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. 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 the vast majority of the entire input.

For test items, we employed the same stimuli used in Author (xxxx). 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 Author (xxxx) involved acoustic masking, the same stimuli of these conditions 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.

Table 3. Composition of test stimuli

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

Model creation and training

With the Python packages and pre-trained models (Table 4), we trained the three NN models with the caregiver input data, along with parameter setting advised by previous studies (e.g., Vázquez et al., 2020; Wu et al., 2019). We used the respective pre-trained models in developing the NN models for the following reasons. NN algorithms typically require a massive amount of training data for their optimal operation. Unfortunately, we are not aware of any 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 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 4. Summary: Model specification

	LSTM	BERT	GPT-2
Package	*PyTorch*	*Transformers*	
Pre-trained model	*KoCharElectra-Base* (Park, 2020); 11,568 syllable types	*KoBERT* (Jeon et al., 2019); 54-million-word tokens	*KoGPT2-base-v2* (Jeon et al., 2019); 51,200-word tokens
Tokenisation	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	

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 (Agent-First; Theme-First) 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 (1 = Agent-First; 0 = 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 separate column 'Label' 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. Model training proceeded with the initial values of epsilon, learning rate, and seed; these values 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 outcomes from 30 times of learning.

The GPT-2 model's training process was almost the same as above, except that the BERT model used particular symbols to mark the start/end of each input sentence, and no such manipulation occurred in the GPT-2 model training. 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 Author (xxxx), on the four case-marked conditions. For the two-argument conditions ($N_{NOM}N_{ACC}V$; $^{\dagger}N_{ACC}N_{NOM}V$), 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 ($N_{NOM}V$; $N_{ACC}V$), all 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.

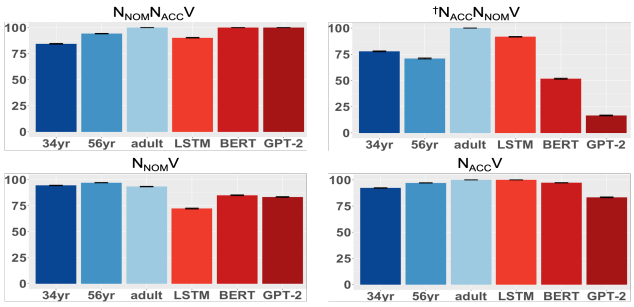

Figure 1. Child comprehension and model performance. X-axis: group (for experiment) or model (for simulation). Y-axis: accuracy. Error bars indicate 95% CI.

These results suggest two major traits 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. There are two important properties of the caregiver input (see Table 2). One is that the number of first-noun-as-agent patterns (3,049 instances) did not exceed that of first-noun-as-theme patterns (3,579 instances). The other property is that 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 (Authors, xxxx). 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 (see Section 3), 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 Author (xxxx), 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).

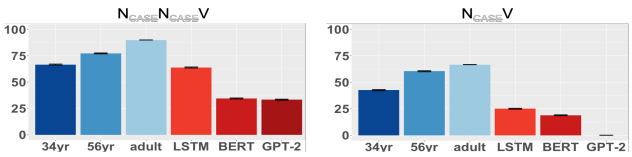

Figure 2. Child comprehension and model performance. X-axis: group (for experiment) or model (for simulation). Y-axis: proportion of agent-first interpretation. Error bars indicate 95% CI.

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$_{\text{CASE}}$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$_{\text{CASE}}$V, the LSTM model improved its performance towards Agent-First when additional information (a second nominal) appeared in N$_{\text{CASE}}$N$_{\text{CASE}}$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.

Discussions

This study's results can be attributed to various factors. To illustrate, the simulation environment in this study may not have perfectly conformed to the experimental setting of Author (xxxx) to the extent that the models utilised relevant information from the stimuli in the exact same way as the children did in the experiment. While the experimental stimuli in Author (xxxx) employed acoustic masking to obscure the case markers so the children would notice that there was something but hidden, the same stimuli types 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 (cf. Stoyneshka et al., 2010). We also used the pre-trained models involving mature language in various genres when constructing each neural-network model, with the consideration of the inherent algorithmic characteristics of neural networks and the actual linguistic environment with which children are surrounded. Together, although we conducted the simulation work as consistently with the experimental setting in Author (xxxx) 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 do highlight that, because these issues have not been fully explored yet in child language research, 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. In addition to the general nature of caregiver input as short/simple utterances and repetition (e.g., Cameron-Faulkner et al., 2003), the NN models may have been affected by the specific word order and/or the presence of case markers in conducting the classification, as shown with †N$_{\text{ACC}}$N$_{\text{NOM}}$V and the two case-less conditions. This aligns with previous reports on language-specific challenges for automatic processing of Korean (e.g., Kim & Ock, 2015; Kim et al., 2007). Since we are not aware of any study on the contribution of language-specific properties to the NNs' performance on child language, this claim awaits further examination.

We also 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. NNs rely heavily on form itself, and this yields a context in a computational sense; however, it is a qualitatively different context from a genuinely linguistic one, 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$_{\text{CASE}}$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$_{\text{NOM}}$V and N$_{\text{ACC}}$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 algorithmic operation differs from how a human processor deals with linguistic knowledge, 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; Özge et al., 2019), despite the same pursuit of efficiency in information processing like a computation model. Given the developing nature of a child processor (e.g., Omaki & Lidz, 2015; Snedeker & Trueswell, 2004; Trueswell et al., 2012), the children in Author (xxxx) 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. We believe that the implications of this study provide a nuanced understanding of the assumed strength of NNs for revealing human language behaviour.

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.

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.

Elman, J. L. (1990). Finding structure in time. *Cognitive science, 14*(2), 179–211.

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.

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.

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.

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.

Kim, W., & Ock, C. (2015). Korean semantic role labeling using case frame and frequency. *The Korean Institute of Information Scientists and Engineers, 6*, 651–653.

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.

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.

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

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

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.

Trueswell, J. C., Kaufman, D., Hafri, A., & Lidz, J. (2012). Development of parsing abilities interacts with grammar learning: Evidence from Tagalog and Kannada. In A. K. Biller, E. Y. Chung & A. E. Kimball (Eds.), *Proceedings of the 36th annual Boston University Conference on Language Development* (pp. 620–632). Somerville, MA: Cascadilla 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.