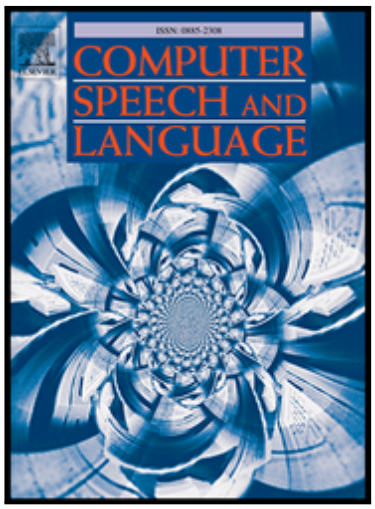


Modelling child comprehension: A case of suffixal passive construction in Korean

Gyu-Ho Shin , Seongmin Mun

PII: S0885-2308(24)00084-6
DOI: https://doi.org/10.1016/j.csl.2024.101701
Reference: YCSLA 101701

To appear in: *Computer Speech & Language*

Received date: 7 January 2024
Revised date: 28 June 2024
Accepted date: 26 July 2024

Please cite this article as: Gyu-Ho Shin , Seongmin Mun , Modelling child comprehension: A case of suffixal passive construction in Korean, *Computer Speech & Language* (2024), doi: https://doi.org/10.1016/j.csl.2024.101701

Highlights

- We explored how neural network models capture monolingual children's comprehension
- The LSTM and GPT-2 models were fine-tuned via patching and hyperparameter adjustments
- We assessed how the models classify Korean suffixal passive sentences used in Shin (2022a)
- The models did not faithfully replicate the children's response patterns in Shin (2022a)
- Our findings highlight the models' limitations in revealing child language features

Modelling child comprehension: A case of suffixal passive construction in Korean

Abstract

The present study investigates a computational model's ability to capture monolingual children's language behaviour during comprehension in Korean, an understudied language in the field. Specifically, we test whether and how two neural network architectures (LSTM, GPT-2) cope with a suffixal passive construction involving verbal morphology and required interpretive procedures (i.e., revising the mapping between thematic roles and case markers) driven by that morphology. To this end, we fine-tune our models via patching (i.e., pre-trained model + caregiver input) and hyperparameter adjustments, and measure their binary classification performance on the test sentences used in a behavioural study manifesting scrambling and omission of sentential components to varying degrees. We find that, while these models' performance converges with the children's response patterns found in the behavioural study to some extent, the models do not faithfully simulate the children's comprehension behaviour pertaining to the suffixal passive, yielding by-model, by-condition, and by-hyperparameter asymmetries. This points to the limits of the neural networks' capacity to address child language features. The implications of this study invite subsequent inquiries on the extent to which computational models reveal developmental trajectories of child language that have been unveiled through corpus-based or experimental research.

Keywords

Neural network; Classification; Child comprehension; Passive construction; Korean

Authors

Gyu-Ho Shin (Department of Linguistics, University of Illinois at Chicago)

Seongmin Mun (Humanities Research Institute, Ajou University)

1. Introduction

One notable trend in language sciences is to apply computational methods and techniques to pursue linguistic inquiries. This line of research has explored computational models' capacity to simulate human language behaviour (Chang, 2009; Hawkins et al., 2020; Jones & Bergen, 2024; Marvin & Linzen, 2019; Warstadt et al., 2019; Wilcox et al., 2018), together with performance-wise variations across algorithms (Hu et al., 2020; Shin & Mun, 2023a), thereby gaining momentum in addressing how learning occurs in the human mind without presuming innate knowledge about grammar (Contreras Kallens et al., 2023; Perfors et al., 2011; Shin, 2021; Shin & Mun, 2023b; Warstadt & Bowman, 2020; but see Perkins et al., 2022). Despite its significance, the current research practice in this field bears three grave limitations. First, the field is skewed heavily towards a limited range of languages (and especially English) and usage features (e.g., adult language). In particular, based on the predominance of English-oriented Large Language Models (LLMs), the intensification of this research bias is being accelerated. This restricts the generalisability of findings from previous studies to lesser-studied languages and registers. Second, while the vast majority of work on this topic seeks to propose new models or improve currently available models, researchers pay relatively little attention to whether and how the implications of computational simulations are compatible with those of other types of measurement, such as behavioural experiments and corpus findings revealing fundamental architectures of human language behaviour. We are aware of few studies informative in this regard (Ambridge et al., 2020; Oh et al., 2022; Shin & Mun, 2023a, 2023b; Xu et al., 2023). This gap prevents explainable AI, that is, an evaluation of the degree to which the performance of computational models addresses the emergence, growth, and change in linguistic knowledge in a sensible, interpretable way. Third, researchers' access to computing resources in academia is limited. Researchers in academia often confront costly access to cutting-edge algorithms and pre-trained models, as well as weak computing power.

These circumstances stifle AI literacy, namely, researchers' ability to understand how computational algorithms work and utilise them to pursue linguistic inquiries. Together, these limitations pose a serious threat to diversity, equity, and inclusion in research (Benders et al., 2021; Blasi et al., 2022; Chang & Bergen, 2024).

The present study aims to alleviate these concerns by investigating how computational models capture children's language behaviour during comprehension, a process in which language users identify an intended meaning or function from a given linguistic form (Goldberg, 2019). In this study, we attend to children as the target population. Despite being extensively investigated in the language acquisition literature due to their notable systematicity and variability of linguistic development interfacing with domain-general learning capacities, this population has remained understudied in computational approaches to language science. With technological advancements, computational methods hold the potential to complement and advance traditional research paradigms by uncovering patterns and mechanisms of child language development, as a type of general knowledge formation process. We specifically focus on neural networks as an artefact of biological neurons in the human brain. To this end, we employ a suffixal passive construction in Korean, which is an understudied language for this topic and is computationally challenging due to its language-specific properties. Cross-linguistically, a passive construction is one major clausal type that expresses a transitive event ('who does what to whom') and poses a challenge to its acquisition for children due to various factors involving the passive voice, such as its paucity of input, the structural complexity that it manifests, and its competition with active-voice knowledge, which is frequent in use and deeply entrenched in the mind (Abbot-Smith et al., 2017; Borer & Wexler, 1987; Brooks & Tomasello, 1999; Huang et al., 2013; Messenger & Fisher, 2018; Shin, 2022a; Shin & Deen, 2023; see also Deen, 2011).

1.1 Computational modelling of children's linguistic knowledge

An emerging line of research applies computational methods to reveal developmental trajectories of linguistic knowledge measured through children's comprehension or production (Alishahi & Stevenson, 2008; Ambridge et al., 2020; Bannard et al., 2009; Martinez et al., 2023; Sagae, 2021; Yedetore et al., 2023; You et al., 2021). Alishahi and Stevenson (2008) conducted Bayesian simulations on acquiring English verb-argument constructions. They created artificial input as pairs of a sentential frame and the corresponding semantic description of that frame based on caregiver input. These form–meaning pairs were used to train a Bayesian learner that displayed probability distributions involving constructional clusters as learning proceeded. The results showed that, as the quantity of input increased over time, the learner was able to assign higher probabilities to frequently occurring verbs within specific constructions to which they were mapped and generalise this schematic knowledge to a newly attested lexicon. You et al. (2021) investigated whether meaning can be acquired with reference to contextual information (generated by word co-occurrences) and without reference to syntactic structures. They trained Word2Vec models with two different types of speech in English (child-directed speech vs. adult-directed speech) and conducted a discrimination task with causality as a test case. The results showed that the models were able to infer causal meaning from simple co-occurrences of neighbouring words in child-directed speech, indicating that word sequences can allow semantic inference without resort to explicit structural information. Sagae (2021) examined the extent to which neural network models track the change in English-speaking children's language throughout learning, which is measured via language assessment metrics. The study trained an LSTM model with longitudinal language data for 16 children (specifically using morphosyntactic tags in the data) and evaluated the model's classification accuracy, measured by age in months. The results showed that the model generally performed

on par with the baseline metrics (mean length of utterance, developmental sentence score, index of productive syntax), indicating that the model could capture linguistic structures relevant to the assessment of language development.

The case for non-English, underrepresented languages and language-usage contexts is extremely thin. For example, Ambridge et al. (2020) tested how children acquire the ability of productive generalisation which also conforms to usage conventions of their native language. They compared acceptability judgements of sentences describing causation events for five typologically distinctive languages (English; Japanese; Hindi; Hebrew; K'iche') and conducted a series of simulations using a discriminative learning mechanism. The model developed for each language was trained on general-purpose corpora in that language and was not provided with acceptability-rating data. Results showed that the model, whose performance was measured by correlating these ratings with human judgments, exhibited a good level of fit to children's and adults' judgement data, except for the case of K'iche'. The findings of Shin and Mun (2023a) are particularly pertinent to the present study. They expanded upon Shin (2021), which measured comprehension behaviour in Korean monolingual children, focusing on the *Agent-First* strategy. Adopting a series of picture-selection tasks involving active transitive sentences with varying degrees of scrambling and omission of sentential components, the children were asked to choose one of two pictures corresponding to an auditory stimulus describing the target picture. Shin and Mun (2023a) investigated the ability of neural networks to simulate the children's picture-selection performance observed in Shin (2021), by assessing the binary classification (Agent-First; Theme-First) performance of four models (Word2Vec, LSTM, BERT, GPT-2) using the test stimuli from the original study. The results revealed that, while there were some similarities between these models' classification performance and the children's response patterns, the

models' performance did not fully align with the children's utilisation of this strategy. This discrepancy highlights asymmetries both across models and across experimental conditions.

1.2 Acquisition of suffixal passive in Korean

Korean is an agglutinative, Subject–Object–Verb language with overt case-marking via dedicated particles and active use of verbal morphology to indicate grammatical information. Two major clausal constructions deliver transitivity in Korean: active transitive and passive. The canonical active transitive pattern in Korean, when fully marked as in (1a), occurs with a nominative-marked agent, followed by an accusative-marked theme; a verb carries no dedicated active morphology. Korean allows scrambling of sentential components as in (1b) if that reordering (agent–theme → theme–agent in this case) preserves the basic propositional meaning. In addition, omission of sentential components is permitted as in (1c-d) if event participants are clearly identified with no ambiguity arising within the context (Sohn, 1999).

(1) Active transitive

a. Canonical

Mina-ka	Pola-lul	an-ass-ta.
Mina-NOM	Pola-ACC	hug-PST-SE[1]

'Mina hugged Pola.'

b. Scrambled

Pola-lul	Mina-ka	an-ass-ta.
Pola-ACC	Mina-NOM	hug-PST-SE

'Mina hugged Pola.'

c. Omission (case marker)

Mina-ka Pola-~~lul~~ an-ass-ta.

Mina-NOM Pola-~~ACC~~ hug-PST-SE

'Mina hugged Pola.'

d. Omission (argument & case marker)

~~Mina-ka~~ Pola-lul an-ass-ta.

~~Mina-NOM~~ Pola-ACC hug-PST-SE

'(Mina) hugged Pola.'

Pertaining to the passive construction, the passive voice is marked across languages (Haspelmath, 1990; Siewierska, 2013), and its usage frequency in Korean is notably low (in comparison to the use of the active voice; Park, 2021; Shin & Mun, 2023b; Woo, 1997). Of the three types of passive construction (Sohn, 1999), the suffixal passive (which is the most frequent type that children encounter; Shin, 2022a) consists of two arguments, a nominative-marked theme and a dative-marked agent occupying the subject and oblique positions, respectively; a verb carries dedicated passive morphology. While the canonical pattern follows the theme–agent ordering as in (2a), it can be scrambled, yielding the agent–theme ordering as in (2b) with the propositional meaning intact.

(2) Suffixal passive

a. Canonical

Pola-ka Mina-hanthey an-ki-ess-ta.

Pola-NOM Mina-DAT hug-PSV-PST-SE

'Pola was hugged by Mina.'

b. Scrambled

Mina-hanthey Pola-ka an-ki-ess-ta.

Mina-DAT Pola-NOM hug-PSV-PST-SE

'Pola was hugged by Mina.'

Passive morphology, which is one of the four allomorphic variants of suffixes *-i-, -hi-, -li-,* or *-ki-* (Sohn, 1999), serves as a key disambiguation point to identify the structural properties of the suffixal passive sentence, forcing a comprehender to revise the initial analysis prior to that morphology. In Korean, a nominative-marked [+human] argument is likely to be interpreted as an agent, and a dative-marked [+human] argument is likely to be interpreted as a recipient; these interpretations are supported by strong mapping between thematic roles and case markers attested in language use (Kim & Choi, 2004; Shin & Mun, 2023b; Sohn, 1999). Therefore, a plausible way of analysing (2) prior to the verb is that *Pola* acts on/for *Mina*. However, this initial analysis is incongruent with the passive-voice information conveyed by verbal morphology. Thus, upon encountering the verb at the sentence-final position, a comprehender must revise the initial interpretation by recalibrating the arguments' thematic roles as required by passive morphology, mapping a theme role onto the nominative-marked entity and an agent role onto the dative-marked entity. This revision process driven by passive morphology as a late-arriving cue is linguistically and cognitively demanding (Rapp & Kendeou, 2007; Trueswell et al., 1999), thereby adding difficulty in children's comprehension of this construction (Kim et al., 2017; Shin, 2022a; Shin & Deen, 2023).

Shin (2022a), the baseline of the present study, investigated Korean monolingual children's comprehension behaviour involving the suffixal passive construction through four picture-selection experiments combined with a novel methodology that systematically

omitted or obscured portions of test sentences using acoustic sounds (e.g., cough, chewing). In each experiment, a pair of two pictures was presented involving the same action but reversed thematic roles (e.g., a dog kicking a cat; a cat kicking a dog), and a sentence indicating one of the two pictures (e.g., *kangaci-ka koyangi-hanthey cha-i-eyo*. dog-NOM cat-DAT kick-PSV-SE 'The dog is kicked by the cat.') was presented twice orally; participants (three-and-four-year-olds; five-and-six-year-olds; adults) were asked to choose a picture that matched the sentence. The four experiments yielded three key findings concerning children's comprehension of the suffixal passive construction (Table 1). First, given the competition between passive-voice knowledge (induced by verbal morphology) and active-voice knowledge (which is frequent in use and well-entrenched in children's minds), utilising passive-voice knowledge during comprehension was influenced by age (serving as a proxy for language-usage experience). Second, children aged five to six demonstrated the ability to apply passive-voice knowledge, with the degree of its use inversely proportional to the computational complexity of the sentence (e.g., number of arguments, type of case markers present/absent). Third, children aged three and four did not consistently interpret passive sentences in an active-like manner. These findings indicate an emerging sensitivity to passive morphology and a growing capacity to employ passive-voice knowledge tied to that morphology with age, in conjunction with the interplay between voice-related knowledge involving a given stimulus. This suggests early emergence but late mastery of linguistic knowledge, the maturation of which necessitates substantial language-usage experience.

Table 1. Summary of experimental results: Shin (2022a)

Experiment	Condition	Three-and-four-year-olds		Five-and-six-year-olds		Adult	
		Mean	SD	Mean	SD	Mean	SD
1	$N_{NOM}N_{ACC}V_{act}$	0.844	0.36	0.942	0.24	1.000	0.00
	$N_{ACC}N_{NOM}V_{act}$	0.778	0.42	0.710	0.46	1.000	0.00
	$N_{NOM}N_{DAT}V_{psv}$	0.456	0.50	0.478	0.50	1.000	0.00
	$N_{DAT}N_{NOM}V_{psv}$	0.511	0.50	0.768	0.43	1.000	0.00
2	$N_{\sout{CASE}}N_{\sout{CASE}}V_{act}$	0.667	0.48	0.773	0.42	0.900	0.30
	$N_{\sout{CASE}}N_{\sout{CASE}}V_{psv}$	0.545	0.50	0.424	0.50	0.150	0.36
3	$N_{NOM}V_{act}$	0.944	0.23	0.971	0.17	0.933	0.25
	$N_{ACC}V_{act}$	0.922	0.27	0.971	0.17	1.000	0.00
	$N_{NOM}V_{psv}$	0.522	0.50	0.710	0.46	0.967	0.18
	$N_{DAT}V_{psv}$	0.533	0.50	0.841	0.37	0.950	0.22
4	$N_{\sout{CASE}}V_{act}$	0.426	0.50	0.604	0.50	0.667	0.48
	$N_{\sout{CASE}}V_{psv}$	0.593	0.50	0.333	0.48	0.100	0.30

Note. The scoring for the conditions in Experiments 2 and 4, which can in principle be interpreted in more than one way, was based on the high likelihood of agent-first interpretation (0: theme-first; 1: agent-first). The mean scores in these conditions indicate the mean rates of agent-first response.

1.3 The present study

In addition to the acquisitional challenges involving the passive voice as shown across languages, this construction poses an additional challenge to Korean monolingual children because passive morphology in a verb invokes a mandatory revision of initial interpretation on the associations between thematic roles and case markers from typical/frequent (nominative-marked agent; dative-marked recipient) to atypical/infrequent (nominative-marked theme; dative-marked agent) ones. In this respect, great interest lies in whether computational models can recognise passive morphology and properly conduct the required

revision process to arrive at the correct interpretation of a suffixal passive sentence. We investigate this issue by developing neural network models with (i) fine-tuning via patching (i.e., pre-trained model + caregiver input) and (ii) hyperparameter variations and by examining their classification performance on the same test stimuli as that used in Shin (2022a). Caregiver input is noteworthy because of its simple, brief, and repetitive nature, which qualitatively differs from adult-directed speech and plays a substantial role in the way that children develop linguistic knowledge (Behren, 2006; Cameron-Faulkner et al., 2003; Snow, 1972; Stoll et al., 2009). Therefore, it is reasonable to assume that a computationally simulated learner trained on caregiver input would elucidate child language features (Alishahi & Stevenson, 2008; You et al., 2021; but see Yedetore et al., 2023). Reflecting the core assumption of usage-based constructionist approaches—what-you-see-is-what-you-get (Goldberg, 2019; Lieven, 2010; Tomasello, 2003), the models engage only in formal features (i.e., raw text) in the course of training and classification, which differs from other studies implementing additional devices in their simulations, such as thematic role variables (Chang, 2002) and a separate layer encoding semantic information (Alishahi & Stevenson, 2008). This study also builds on Shin and Mun (2023a), complementing how hyperparameter variations modulate model performance with respect to child language data. As we are not aware of any study touching upon this inquiry, our study is pioneering and innovative, and simultaneously, somewhat explorative.

In this study, we adopt two neural network architectures: LSTM (Long Short-Term Memory; Hochreiter & Schmidhuber, 1997) and GPT-2 (Generative Pre-trained Transformer 2; Radford et al., 2019). LSTM is a recurrent neural network algorithm with the addition of three gates (*Forget*, determining whether the incoming information from the previous timestamp is irrelevant and thus forgotten; *Input*, quantifying the significance of new information carried by the incoming input; *Output*, submitting the currently updated

information to the next timestamp) comprising a memory cell in a hidden layer. In addition to the possibility that recurrent neural networks learn some aspects of syntactic structures when provided with appropriate training (Kiperwasser & Goldberg, 2016; Futrell & Levy, 2019; Linzen & Baroni, 2021), this algorithm has a better control for the extent to which information in a hidden state is updated after each word. GPT utilises an attention mechanism for effective computation by enhancing each part of the input sequence in consideration of various information about the whole sequence (e.g., segment position) to better identify the most relevant parts of that sequence (Vaswani et al., 2017). Because this algorithm targets a general-purpose learner whose learning trajectories are not subject to particular tasks, model training does not stand on the specifics of data or tasks (Radford et al., 2019); it can also perform new tasks with a relatively small number of examples. Despite the continuous development of the GPT-*n* series, GPT-2 is often employed to conduct simulations on language behaviour (Goldstein et al., 2022; Hosseini et al., 2022), yielding successful modelling on various language tasks.

2. Methods

Figure 1 presents the entire workflow of computational simulations in this study. All the modelling work was conducted using a MacBook Pro (Apple M2 Max with 38-core GPU, 16-core Neural Engine, 96GB unified memory).

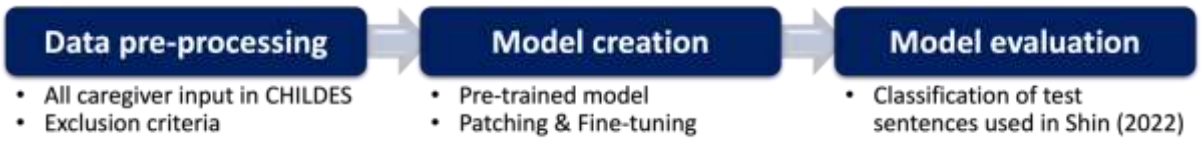


Figure 1. Overview of computational modelling

2.1 Data pre-processing

Table 2 summarises the information about the caregiver-input data in CHILDES (MacWhinney, 2000) used in our study. We utilised the same data as that used by Shin and Mun (2023a) for the current study in consideration of the comparability of findings between the two studies. The data were pre-processed by (i) correcting typos and spacing errors and (ii) excluding any sentence whose length was less than five characters or those consisting only of onomatopoeic and mimetic words (see Shin, 2022b for the details about the pre-processing), which resulted in 69,498 sentences (285,350 eojeols[3]).

Table 2. Information about caregiver input data in CHILDES

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

Note. F = father; GM = grandmother; GF: grandfather; M = mother.

2.2 Model training[4]

2.2.1 Architecture-general procedure

Table 3 provides details on each model created in this study. Neural networks typically require extensive training data for training to ensure their optimal operation (Edwards, 2015), but 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 encounter more than just caregiver input in real-life scenarios; there are many types of exposure to language use with which children are surrounded. To cope with these issues, we employed the respective pre-trained models, which were open-access and representative at

the moment of study, and patched the caregiver-input data to each pre-trained model when developing our models.[5] The patching procedure, inspired by prior work (Ilharco et al., 2022; Moon & Okazaki, 2020; Ninalga, 2023), involved enlarging a pre-trained model by incorporating syllables from the caregiver input which were not present in the model to that model. This procedure increased the vocabulary size of the GPT-2 pre-trained model (51,200 to 67,052). We believe that incorporating caregiver input into pre-trained models can enhance ecological validity for this type of modelling, but no research has scrutinised this point thoroughly, indicating the need for further attention.

Table 3. Specification of computational models

	LSTM	GPT-2
Python Package	*PyTorch* (Paszke et al., 2019; version 2.1.0)	*Transformers* (Wolf et al., 2020; version 4.35)
Pre-trained model	*KoCharElectra-Base*[(a)] (Size: 11,360)	*KoGPT2-base-v2*[(b)] (Size: 51,200)
Tokenisation	Syllable-based	Syllable-based; *Byte Pair Encoding*
Hyperparameter variation	Learning rate: 0.001, 0.0001 Batch size: 16, 32, 64 Dropout rate: 0.3, 0.5, 0.7	Learning rate: 0.001, 0.0001 Batch size: 16, 32, 64 Max. sequence length: 64, 128, 256
Epoch	10	10
Model-specific	Hidden layers: 256 Embedded dimension: 128 Hidden dimension: 8 Number of layers: 1	Seed: 42 Epsilon: 0.00000001 Embedding & hidden dimension: 768 FFN inner hidden dimension: 3,072 Number of attention heads: 12 Number of parameters: 125M Number of transformer layers: 12

Note. (a) https://github.com/monologg/KoCharELECTRA (accessed on 2023-11-07). (b) https://github.com/SKT-AI/KoGPT2 (accessed on 2023-11-07).

To conduct the binary classification of test items (Agent-First; Theme-First), our models were further fine-tuned on 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 (see Appendix for the information about the instances). The instances were extracted from the caregiver-input data in CHILDES through an automatic search process developed by Shin (2022b); every sentence for each extraction was checked manually to confirm its accuracy. This treatment also aimed to enhance compatibility between the simulation environments and the experimental settings of Shin (2022a), in which participants were shown transitive-event pictures before receiving a stimulus to contextualise their interpretation. The procedure involved exposing the models to two labels indicating the thematic-role orderings of transitive-event sentences, along with the sentences themselves, to prepare the models for the designated classification task. This approach is conceptually analogous to the procedure employed with children in Shin (2022a). 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.

To investigate the influence of hyperparameter variations on model performance when handling child language data, we adjusted three hyperparameters for each architecture: learning rate, batch size, and dropout rate for LSTM; learning rate, batch size, and maximum sequence length for GPT-2. Our choices were informed by previous studies (architecture-general: Li et al., 2020; Sun et al., 2019; Takase et al., 2018; Wu et al., 2019; LSTM: Kågebäck & Salomonsson, 2016; Ma et al., 2020; Qian et al., 2017; Yang et al., 2019; GPT-2: Budzianowski & Vulić, 2019; Dai et al., 2023; Oh & Schuler, 2022; de Vries & Nissim, 2021). These variations resulted in 18 sub-models per architecture.

2.2.2 Architecture-specific procedure

2.2.2.1 LSTM

No syllable-based Korean pre-trained model for this architecture exists, so we extracted relevant vocabulary information from a pre-trained model for ELECTRA and trained the model. For each epoch, all the syllable information was submitted to the model's input layer. Take an eojeol *twayci-ka* 'pig-NOM' as an example (Figure 2). For the syllable *ci*, the model first evaluates if the information about the previous syllable *tway* obtained from the prior cell is relevant to the current input at the Forget gate (σ1). The model then quantifies the information about the current input via the tangent function at the Input gate (σ2). Finally, the model hands over this outcome to the processing of the next syllable *ka* at the Output gate (σ3), again via the tangent function. Once a sentence is completed for processing, the optimiser computes the distance/loss between the observed value and the predicted value, the result of which is transmitted through backpropagation.

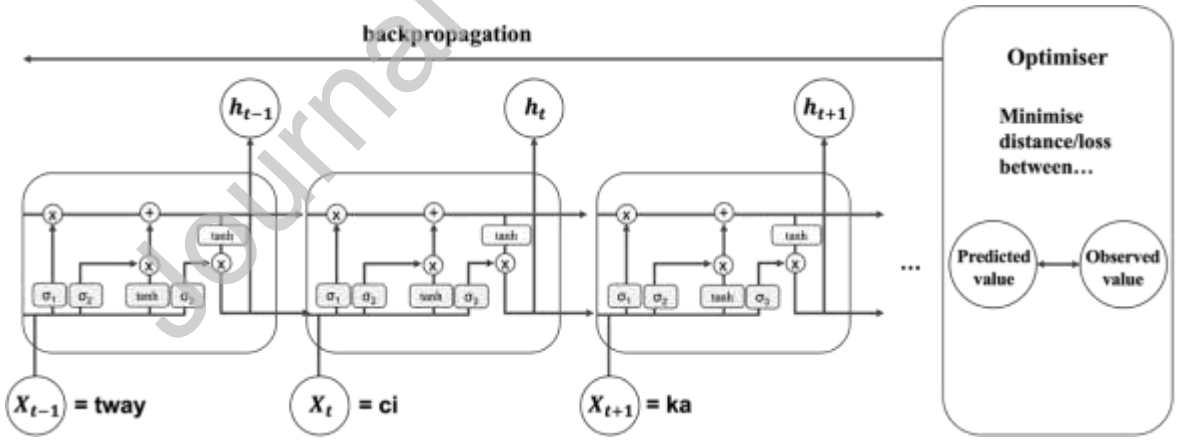


Figure 2. Model training: LSTM (e.g., *twayci-ka* 'pig-NOM')

After the training, the model evaluated the test stimuli, accumulating by-syllable information sequentially (by generating respective hidden layers) and then comparing the outcomes (1 = Agent-First; 0 = Theme-First) to the actual labels of these stimuli. We repeated the same

learning process 30 times in each epoch and averaged the by-condition outcomes in assessing the models' classification performance to alleviate potential variations during the task.

2.2.2.2 GPT-2

As illustrated in Figure 3, each input sentence in the fine-tuning stage was transformed into two embedding types. For token embedding, the sentences were tokenised as syllable units. Originally, GPT-2 utilised a character for this task in the case of English. However, KoGPT-2 employs a syllable as a basic unit of tokenisation, likely in consideration of the language-specific properties of Korean. For position embedding, each token was converted into a numeric value indicating a unique index of the token with reference to the vocabulary in the patched pre-trained model. The maximum dimension size of position embeddings was determined by the maximum sequence length set in the hyperparameter-setting stage. The initial values of epsilon (i.e., the upper bound of randomness for a model to explore the data) and seed (i.e., the initialisation state of a pseudo-random number generator indicating where a model starts) were automatically updated with the outcomes of each epoch. The training occurred from the initial model with the zero value of gradients to an optimal model with updated values through feedforward and backpropagation. Finally, the trained model per epoch classified the test stimuli; likewise for the LSTM model, we averaged the by-condition classification outcomes from 30 times of learning.

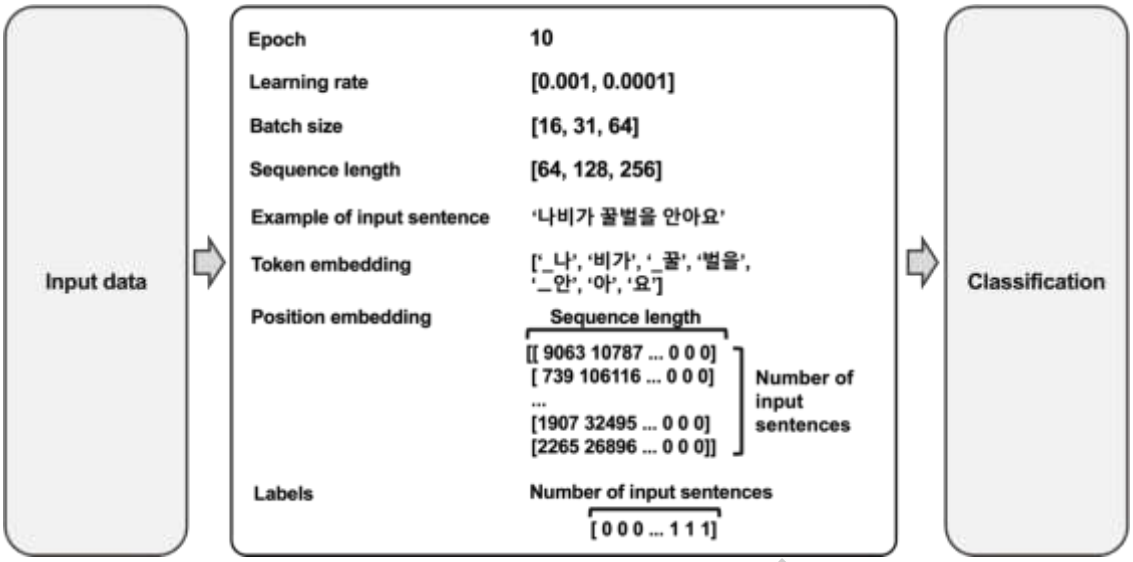


Figure 3. Model training: GPT-2 (e.g., *napi-ka kkwulpel-ul an-ayo* butterfly-NOM honeybee-ACC hug-SE 'The butterfly hugs the honeybee.')

2.3 Model evaluation

For test items, we employed the same stimuli used in Shin (2022a). Each condition consisted of six instances, with animals as agents and themes and actional verbs at the end, as illustrated in Table 4. Each trained model classified every test stimulus, evaluating whether the stimulus fell into Agent-First or Theme-First. While the stimuli in the case-less conditions ($N_{\not CASE}N_{\not CASE}V_{act}$, $N_{\not CASE}N_{\not CASE}V_{psv}$, $N_{\not CASE}V_{act}$, $N_{\not CASE}V_{psv}$) in Shin (2022a) involved acoustic masking effects, the same stimuli in the simulations did not have such auditory effects. This was unavoidable considering this study's simulation setting, in which the models worked exclusively with the textual data. We concede that this difference may serve as one confounding factor for interpreting the results. In this regard, using a [MASK] token, although still textual, may pave the way for further research based on this study's findings.

Table 4. Composition of test stimuli

Condition	Example	Expected classification
$N_{NOM}N_{ACC}V_{act}$	cat-NOM dog-ACC kick	Agent-first
$N_{ACC}N_{NOM}V_{act}$	dog-ACC cat-NOM kick	Theme-first
$N_{NOM}N_{DAT}V_{psv}$	cat-NOM dog-DAT kick-PSV	Theme-first
$N_{DAT}N_{NOM}V_{psv}$	dog-DAT cat-NOM kick-PSV	Agent-first
$N_{\text{~~CASE~~}}N_{\text{~~CASE~~}}V_{act}$[(a)]	cat dog kick	Agent-first
$N_{\text{~~CASE~~}}N_{\text{~~CASE~~}}V_{psv}$[(a)]	cat dog kick-PSV	Theme-first
$N_{NOM}V_{act}$	cat-NOM kick	Agent-first
$N_{ACC}V_{act}$	dog-ACC kick	Theme-first
$N_{NOM}V_{psv}$	cat-NOM kick-PSV	Theme-first
$N_{DAT}V_{psv}$	dog-DAT kick-PSV	Agent-first
$N_{\text{~~CASE~~}}V_{act}$[(a)]	dog kick	Agent-first
$N_{\text{~~CASE~~}}V_{psv}$[(a)]	dog kick-PSV	Theme-first

Note. As (a) can in principle be interpreted in more than one way, the expected classification was determined on the basis of the canonical thematic-role ordering in each construction type (active transitive: Agent-first [agent–theme]; suffixal passive: Theme-first [theme–agent]).

Our aim was to compare the picture-selection performance of children as observed in Shin (2022a) directly and meaningfully with the classification performance of the models in our study. To achieve this, we utilised the models' classification accuracy (or their Agent-First classification rate for the case-less conditions) as analogous to the children's response patterns in each condition. We note that 50%, or a value of 0.5, represents the chance level when interpreting the results.

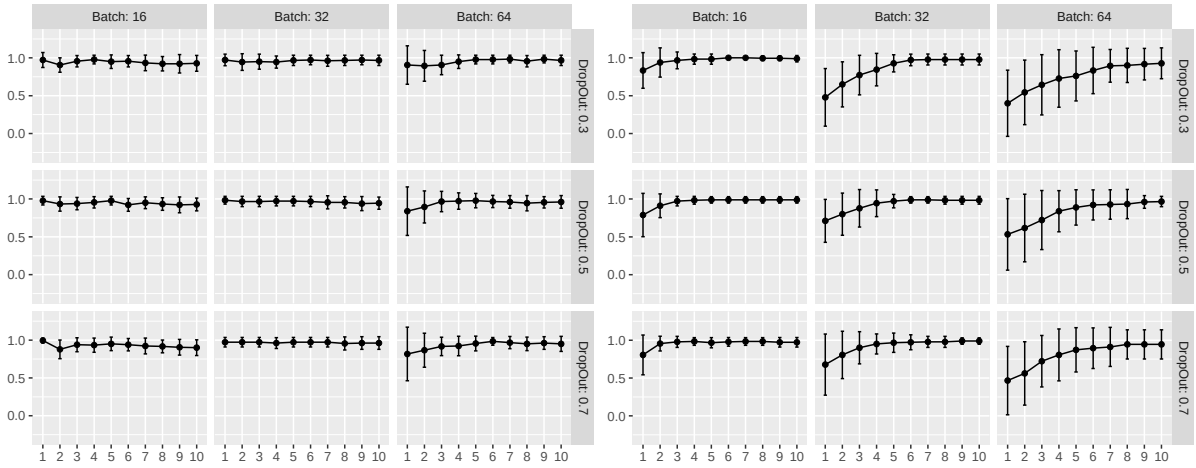
3. Results

3.1 Case-marked conditions

3.1.1 Two-argument active transitive: $N_{NOM}N_{ACC}V_{act}$ & $N_{ACC}N_{NOM}V_{act}$

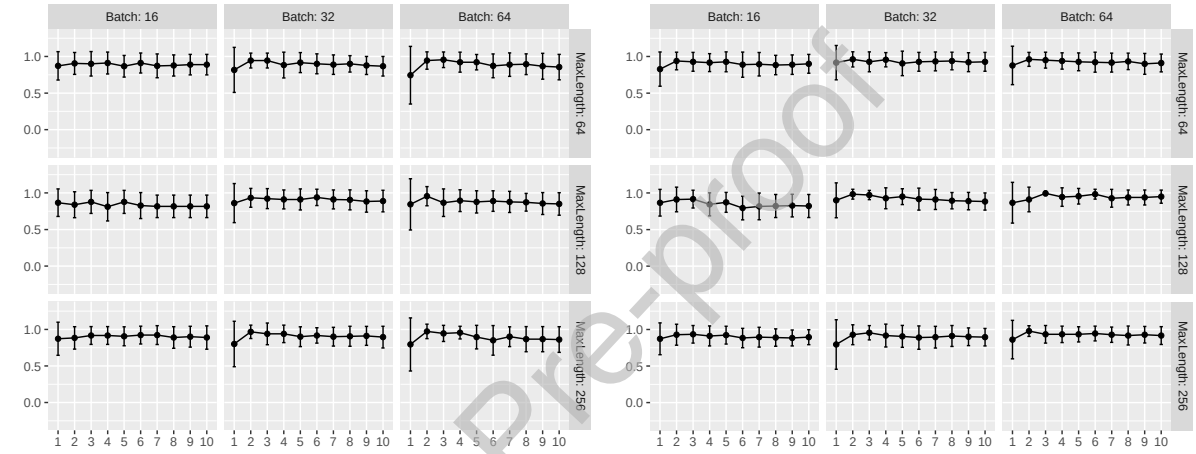
The children in Shin (2022a) were good at both conditions in general, and they were better in the canonical condition than the scrambled condition (three-and-four-year-olds: 84% in $N_{NOM}N_{ACC}V_{act}$ & 78% in $N_{ACC}N_{NOM}V_{act}$; five-and-six-year-olds: 94% in $N_{NOM}N_{ACC}V_{act}$ & 71% in $N_{ACC}N_{NOM}V_{act}$). These findings align with those of previous research showing children's degraded accuracy rates for the scrambled word order relative to the canonical word order (e.g., Jin et al., 2015; Kim et al., 2017; Schipke et al., 2012).

Figures 4 and 5 present the classification accuracy of the models per epoch in each condition. In $N_{NOM}N_{ACC}V_{act}$, all the models demonstrated high accuracy, independently of architecture or hyperparameter types, as the epoch progressed. In contrast, the two architectures showed distinctive performance in $N_{ACC}N_{NOM}V_{act}$: while the LSTM models achieved very high accuracy, the GPT-2 models' accuracy rates were close to 0, regardless of architecture or hyperparameter type. This outcome indicates that the GPT-2 models classified the test stimuli in this condition into Agent-First most of the time (which should have been Theme-First). These findings resemble those of Shin and Mun (2023a), showing the two models' contrastive performance in this condition.



(a) LSTM, learning rate = 0.001

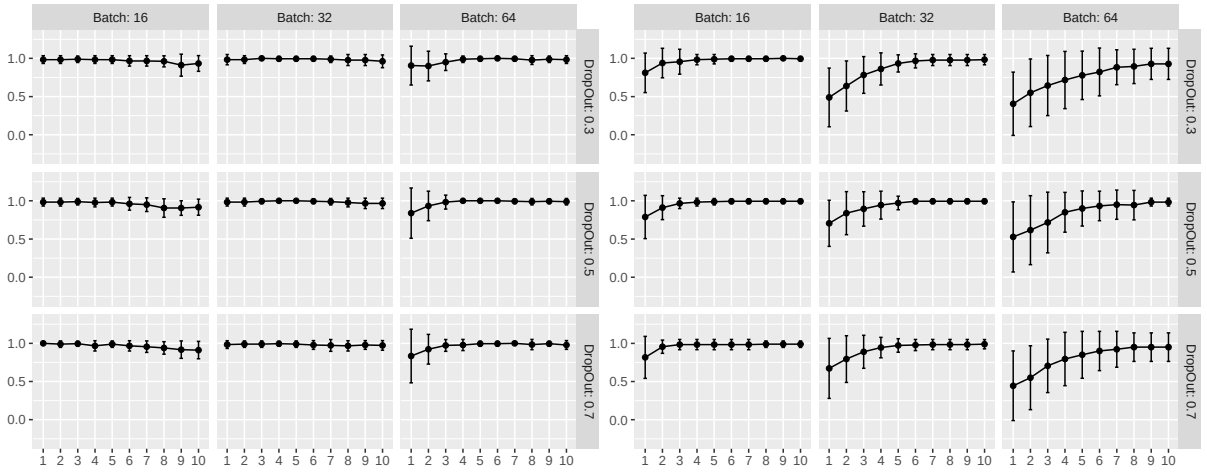
(b) LSTM, learning rate = 0.0001



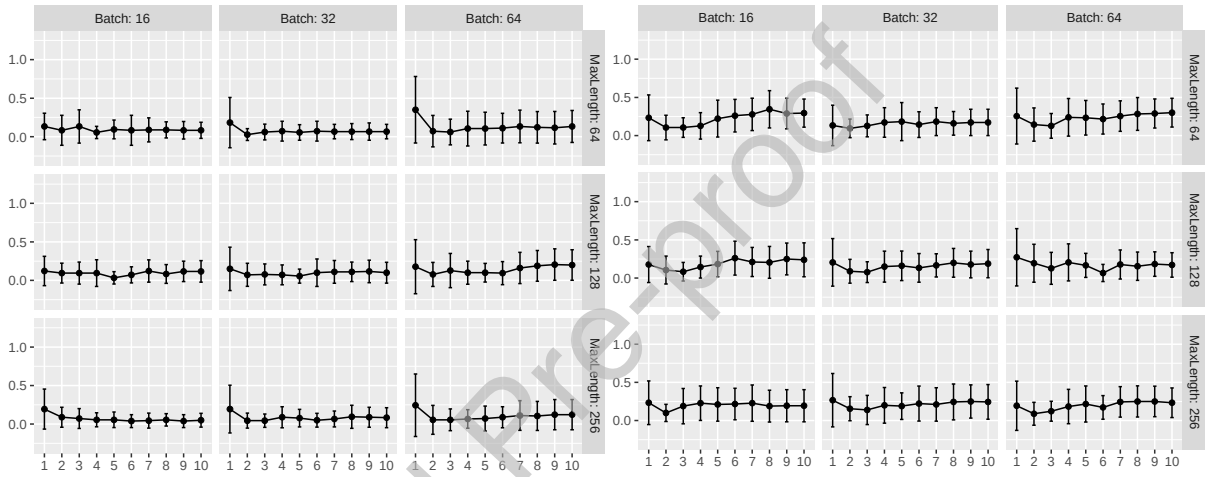
(c) GPT-2, learning rate = 0.001

(d) GPT-2, learning rate = 0.0001

Figure 4. Model performance: $N_{NOM}N_{ACC}V_{act}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.



(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001



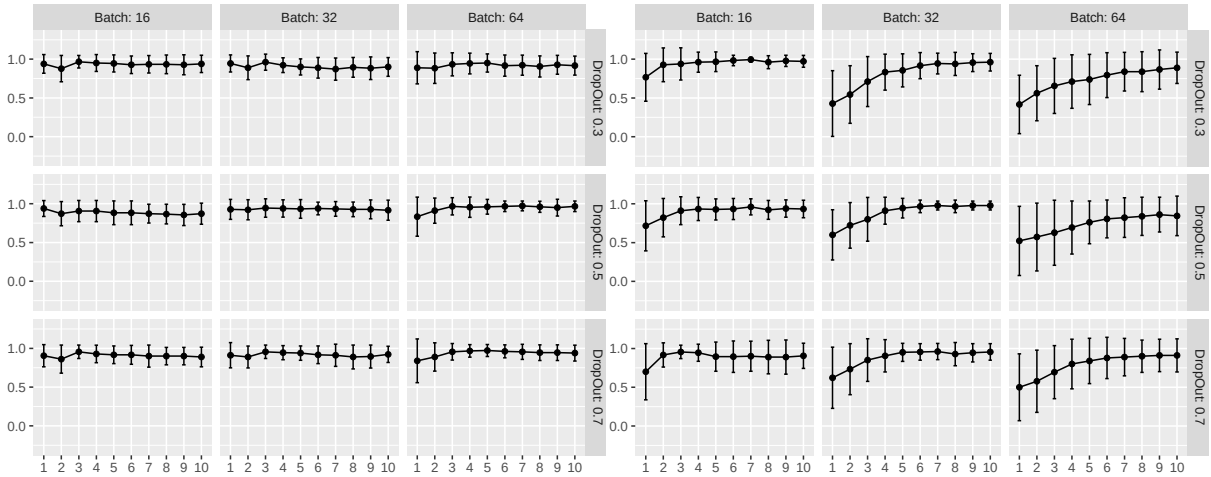
(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 5. Model performance: $N_{ACC}N_{NOM}V_{act}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.

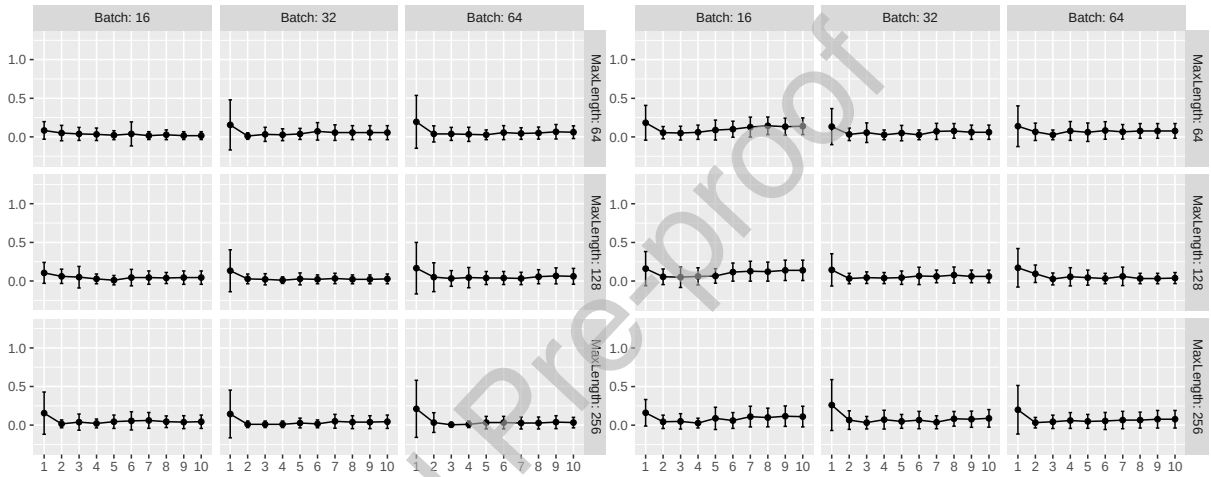
3.1.2 Two-argument suffixal passive: $N_{NOM}N_{DAT}V_{psv}$ & $N_{DAT}N_{NOM}V_{psv}$

The children in Shin (2022a) demonstrated notable by-age-group and by-condition asymmetries when coping with the two passive-voice conditions. While the three-and-four-year-olds showed at-chance performance in both conditions (46% in $N_{NOM}N_{DAT}V_{psv}$; 48% in $N_{DAT}N_{NOM}V_{psv}$), the five-and-six-year-olds showed at-chance performance (51% in $N_{NOM}N_{DAT}V_{psv}$) in the canonical condition and above-chance performance (77% in $N_{DAT}N_{NOM}V_{psv}$) performance in the scrambled condition. These findings indicate that, given the acquisitional challenges involving the passive voice, the children may have noticed passive morphology and utilised passive-voice knowledge tied to that morphology—albeit weak and inconsistent—to some extent, especially in the scrambled condition for the five-and-six-year-olds, against the co-activation of and strong interference from active-voice knowledge.

Figures 6 and 7 present the classification accuracy of the models per epoch in each condition. In $N_{NOM}N_{DAT}V_{psv}$, the two architectures showed distinctive performance: while all the LSTM models achieved very high accuracy, all the GPT-2 models' accuracy rates were close to 0, regardless of architecture or hyperparameter type. This finding indicates that the GPT-2 models predominantly classified the test stimuli in this condition as Agent-First (which should have been Theme-First, the correct interpretation of this condition). However, in $N_{DAT}N_{NOM}V_{psv}$, all the models demonstrated high accuracy, independently of architecture or hyperparameter types, as the epoch progressed.

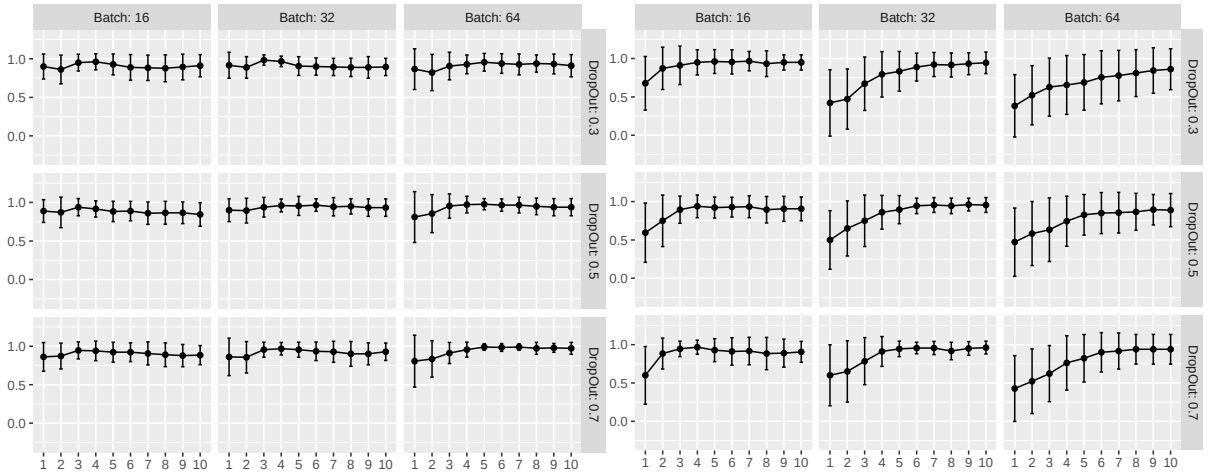


(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001

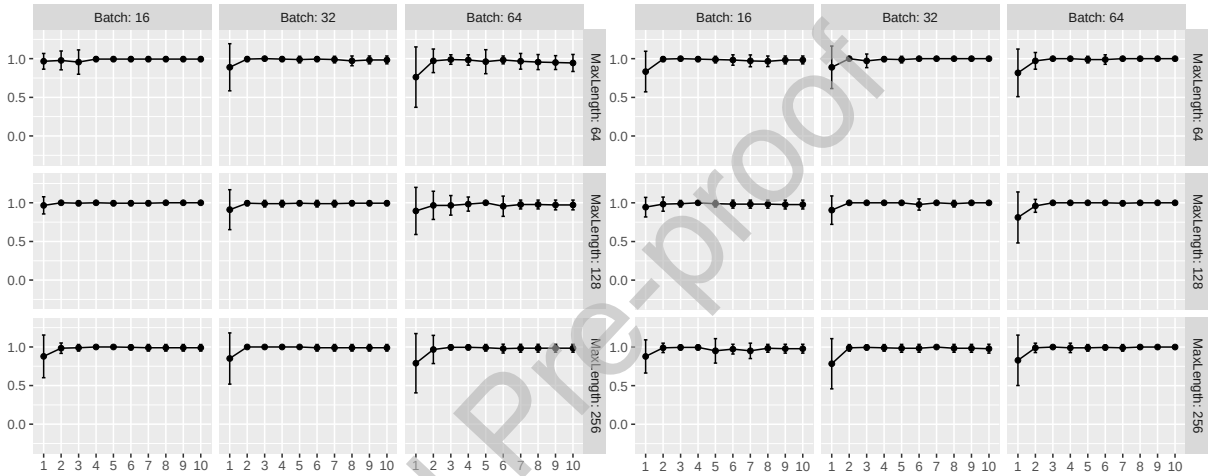


(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 6. Model performance: $N_{NOM}N_{DAT}V_{psv}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.



(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001



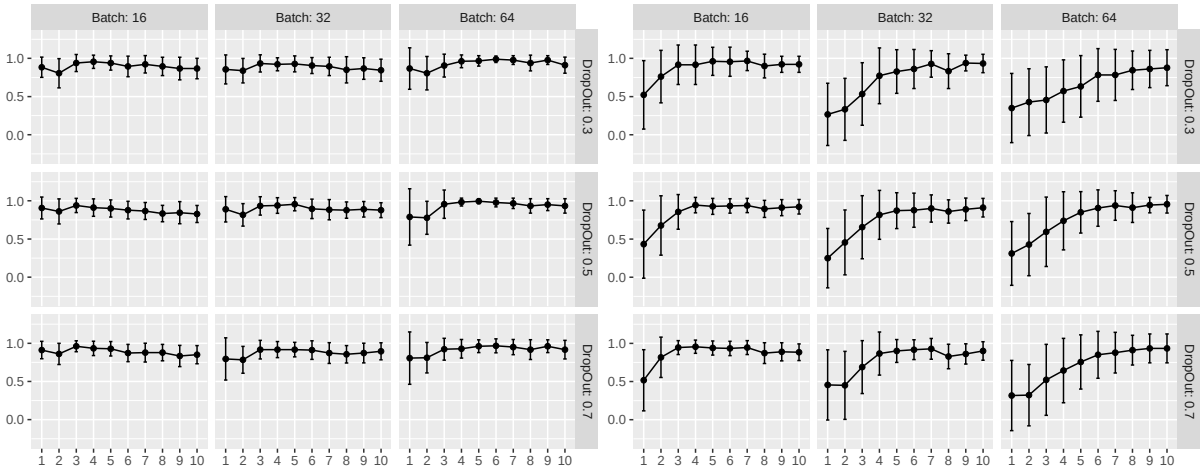
(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 7. Model performance: $N_{DAT}N_{NOM}V_{psv}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.

3.1.3 One-argument active transitive: $N_{NOM}V_{act}$ & $N_{ACC}V_{act}$

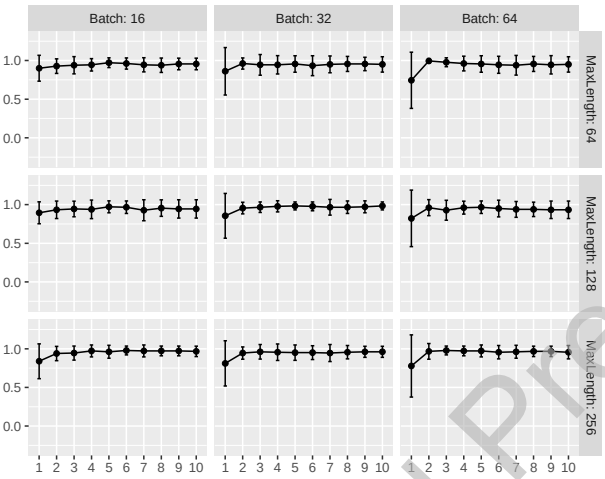
The children in Shin (2022a) were very good at the two conditions (three-and-four-year-olds: 94% in $N_{NOM}V_{act}$ & 92% in $N_{ACC}V_{act}$; five-and-six-year-olds: 97% in both conditions). This finding indicates that they had a good command of the case-marking knowledge required for the active transitive, which is consistent with previous reports (Jin et al., 2015; Özge et al., 2019).

Figures 8 and 9 present the classification accuracy of the models per epoch in each condition. In $N_{NOM}V_{act}$, all the models demonstrated high accuracy, independently of architecture or hyperparameter types, as the epoch progressed. In $N_{ACC}V_{act}$, except for the LSTM models with a learning rate of 0.001, all the models demonstrated high accuracy, regardless of architecture or hyperparameter types, as the epoch progressed. The extraordinary performance of the LSTM models with a learning rate of 0.001 is inconsistent with Shin and Mun's findings (2023a).

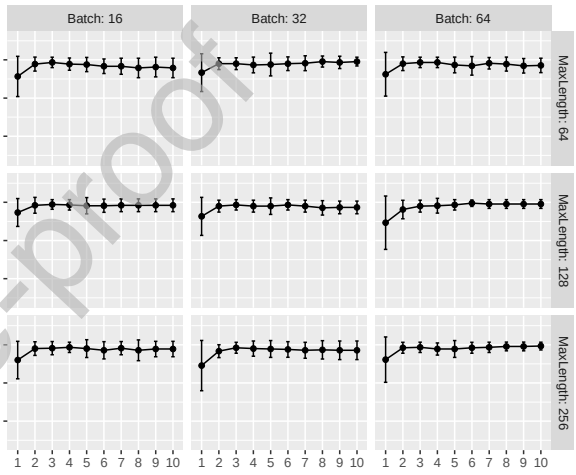


(a) LSTM, learning rate = 0.001

(b) LSTM, learning rate = 0.0001

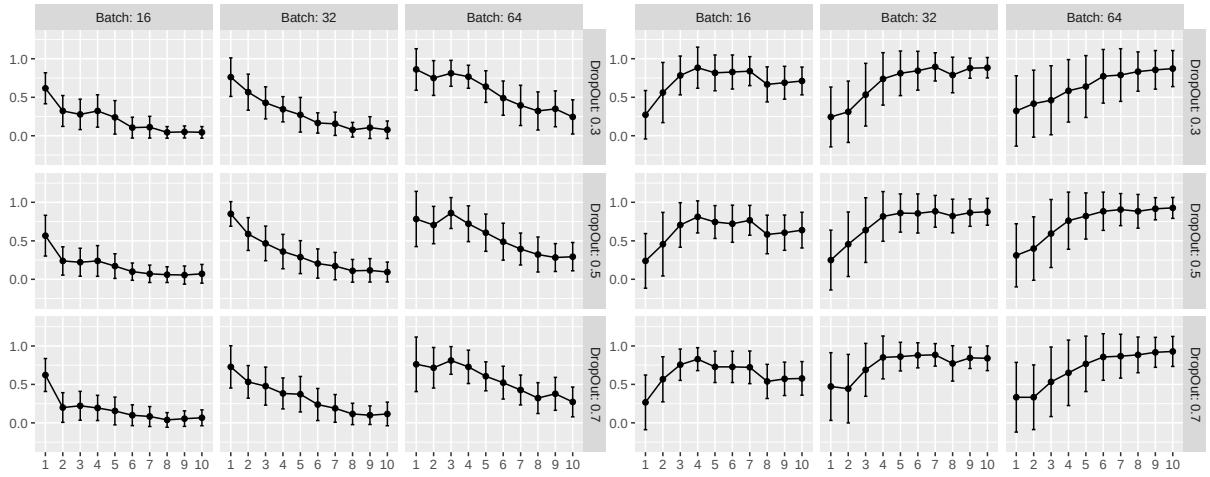


(c) GPT-2, learning rate = 0.001

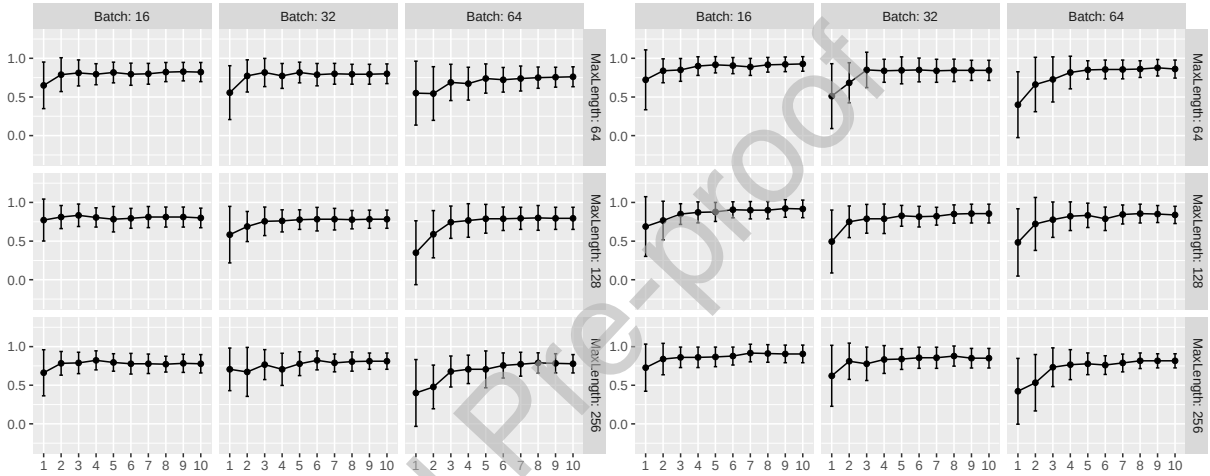


(d) GPT-2, learning rate = 0.0001

Figure 8. Model performance: $N_{NOM}V_{act}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.



(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001



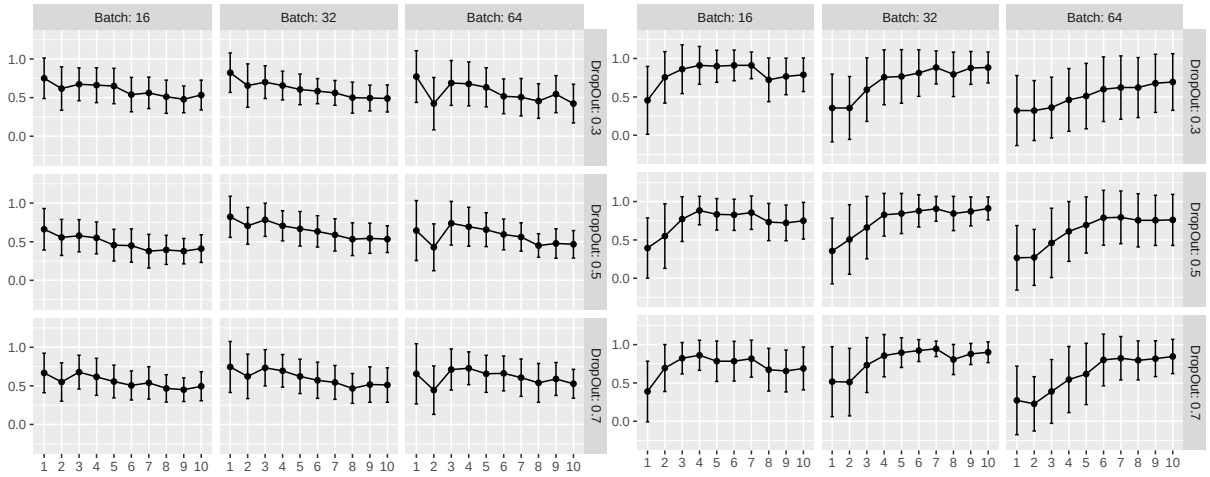
(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 9. Model performance: $N_{ACC}V_{act}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.

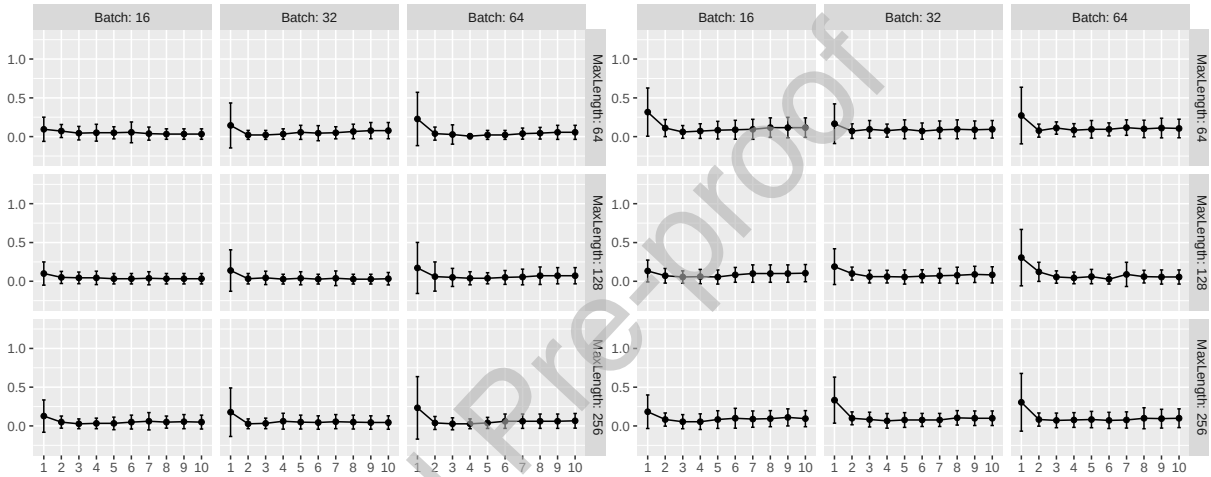
3.1.4 One-argument suffixal passive: $N_{NOM}V_{psv}$ & $N_{DAT}V_{psv}$

The children in Shin (2022a) demonstrated by-age-group differences in accuracy. While the three-and-four-year-olds showed uniformly at-chance performance in the two conditions (52% in $N_{NOM}V_{psv}$; 53% in $N_{DAT}V_{psv}$), the five-and-six-year-olds showed uniformly above-chance performance in both conditions (71% in $N_{NOM}V_{psv}$; 84% in $N_{DAT}V_{psv}$). This finding indicates that passive-voice knowledge may have been increasingly used for sentence comprehension as age increased.

Figures 10 and 11 present the classification accuracy of the models per epoch in each condition. In $N_{NOM}V_{psv}$, the two architectures demonstrated different patterns of accuracy. For LSTM, as the epoch progressed, the models with a learning rate of 0.001 showed at-chance performance, and the models with a learning rate of 0.0001 improved the accuracy up to above-chance performance. For GPT-2, all the models showed very low accuracy, regardless of hyperparameter type, indicating that they classified the test stimuli in this condition into Agent-First most of the time (which should have been Theme-First, the correct interpretation of this condition). In $N_{DAT}V_{psv}$, except the GPT-2 models with the learning rate of 0.0001, all the models demonstrated high accuracy, independently of architecture or hyperparameter types, as the epoch progressed.

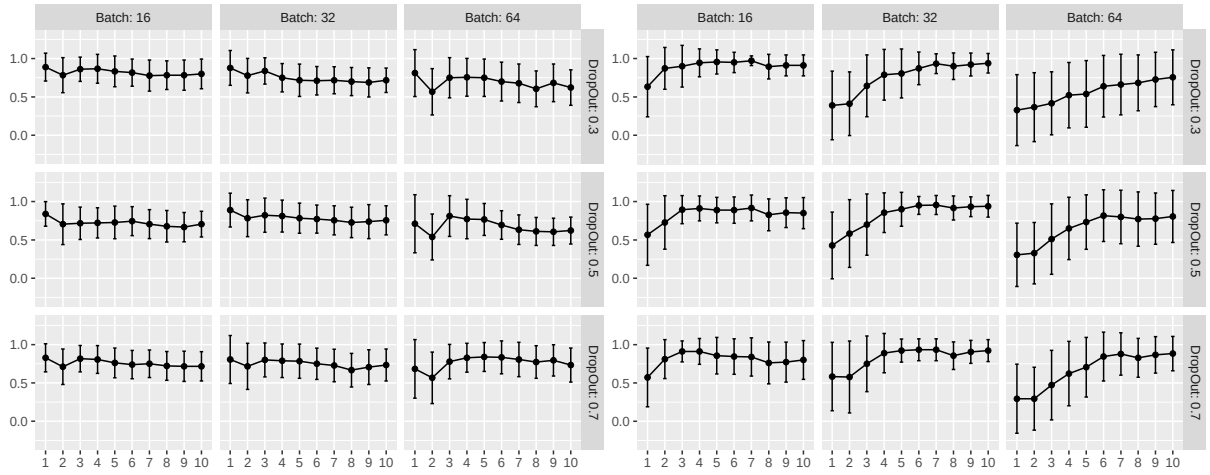


(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001



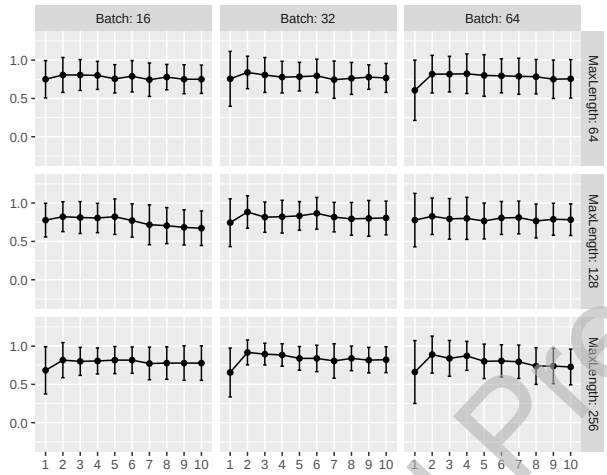
(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 10. Model performance: $N_{NOM}V_{psv}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.

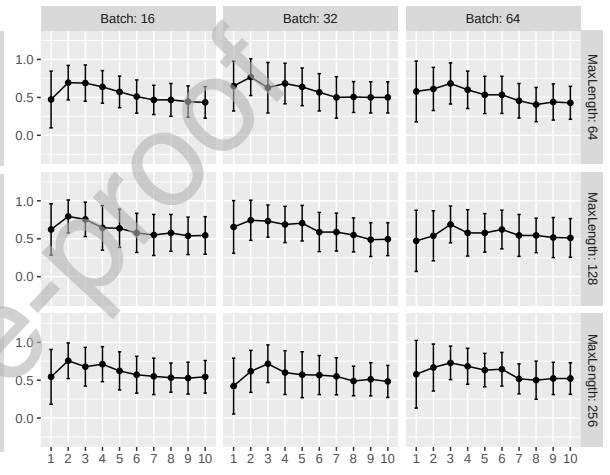


(a) LSTM, learning rate = 0.001

(b) LSTM, learning rate = 0.0001



(c) GPT-2, learning rate = 0.001



(d) GPT-2, learning rate = 0.0001

Figure 11. Model performance: $N_{DAT}V_{psv}$. X-axis = epoch; Y-axis = accuracy (mean). Error bars = 95% CIs.

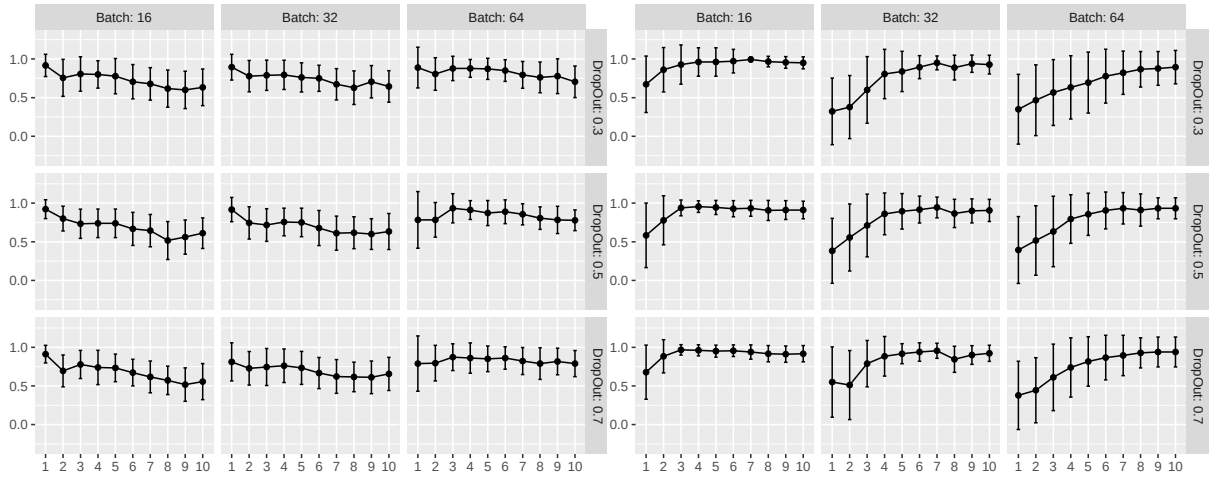
3.2 Case-less conditions

3.2.1 Two-argument conditions: $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{act}$ & $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$

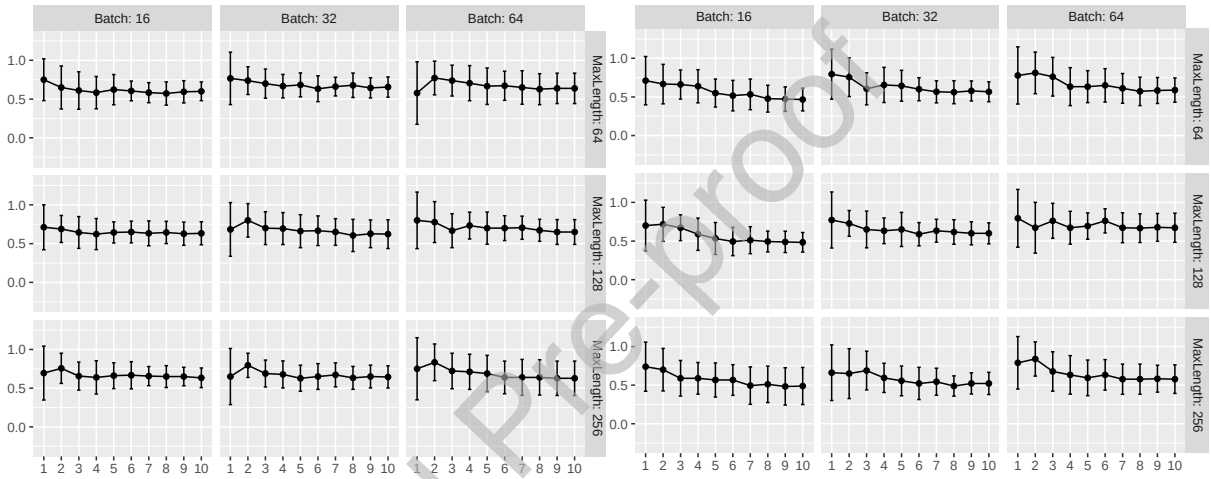
The children in Shin (2022a) showed above-chance performance in $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{act}$, with the five-and-six-year-olds (77%) manifesting more agent-first interpretation than the three-and-four-year-olds (67%). In contrast, they showed numerically lower preference for the agent-first interpretation in $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$ than its active counterpart (54% for the three-and-four-year-olds; 42% for the five-and-six-year-olds), and the difference in the response rates between the two conditions was substantial only for the five-and-six-year-olds. This finding indicates the role of passive morphology in the children's interpretations, with age effects on applying passive-voice knowledge to sentence comprehension. The adult controls' agent-first response rate in $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$ was only 15 per cent, indicating a strong theme-first interpretation in this condition.

Figures 12 and 13 present the classification performance (coded as Agent-First = 1) of the models per epoch in each condition. Overall, the two architectures demonstrated different patterns of classification as the epoch progressed. In $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{act}$, whereas only some of the LSTM models with a learning rate of 0.001 (batch = 64, dropout = 0.5 or 0.7) achieved above-chance rates of Agent-First, all the LSTM models with a learning rate of 0.0001 showed a very high rate of Agent-First. In contrast, all the GPT-2 models were at-chance or slightly above at-chance, regardless of hyperparameter types. The two architectures' performance in this condition aligns partially with Shin and Mun (2023a). A similar kind of by-architecture divergence occurred in $N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$. For LSTM, all the models with a learning rate of 0.001 showed below-chance performance, indicating that they classified the test stimuli in this condition into Theme-First most of the time; all the models with a learning rate of 0.0001 showed above-chance performance, indicating that they predominantly classified the test stimuli in this condition into Agent-First (which should have been Theme-

First, the preferred interpretation of this condition). All the GPT-2 models were at-chance or slightly above-chance, regardless of hyperparameter types.

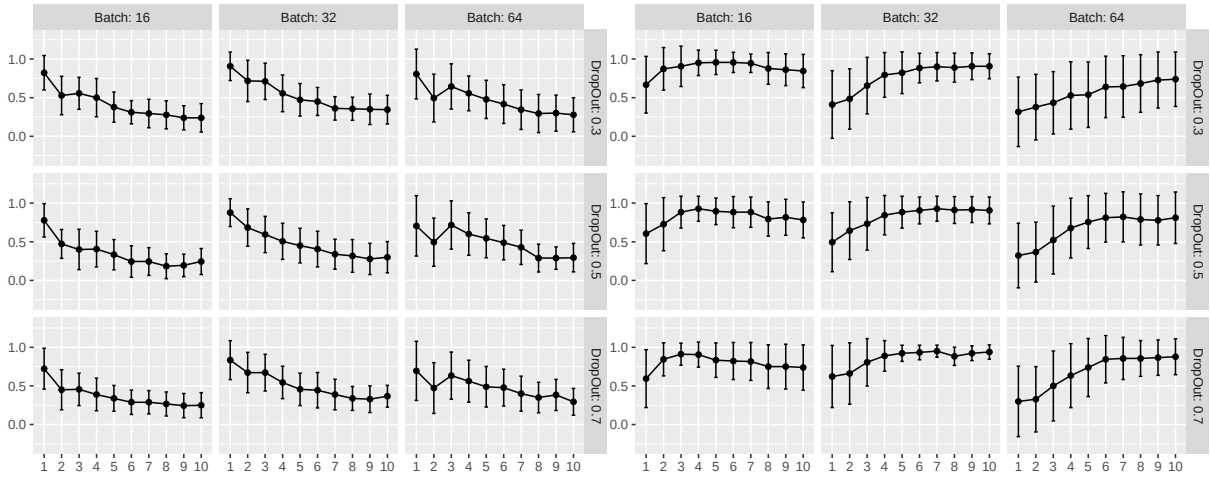


(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001

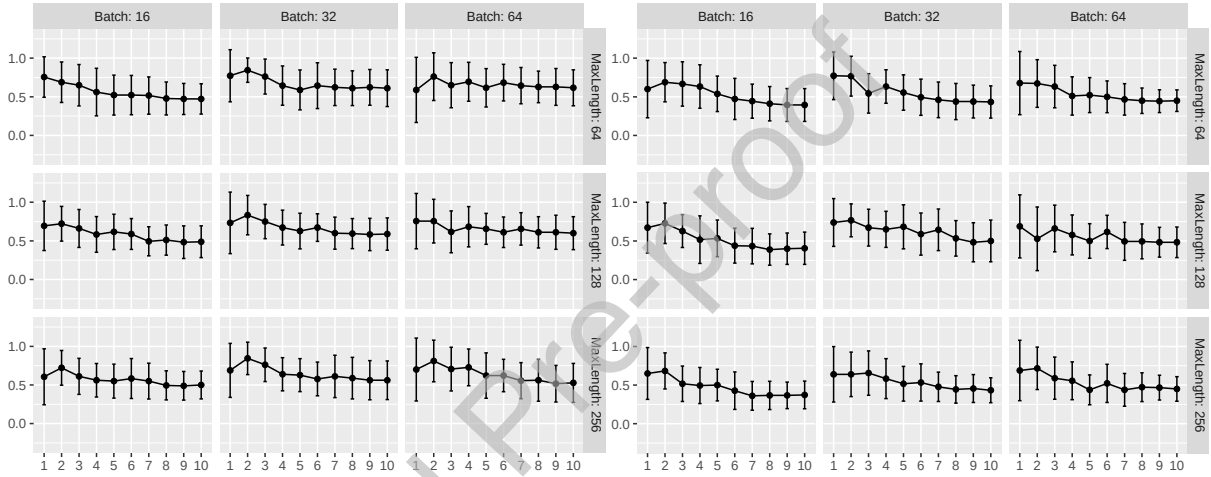


(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 12. Model performance: $N_{CASE}N_{CASE}V_{act}$. X-axis = epoch; Y-axis = agent-first rate (mean). Error bars = 95% CIs.



(a) LSTM, learning rate = 0.001 (b) LSTM, learning rate = 0.0001



(c) GPT-2, learning rate = 0.001 (d) GPT-2, learning rate = 0.0001

Figure 13. Model performance: $N_{CASE}N_{CASE}V_{psv}$. X-axis = epoch; Y-axis = agent-first rate (mean). Error bars = 95% CIs.

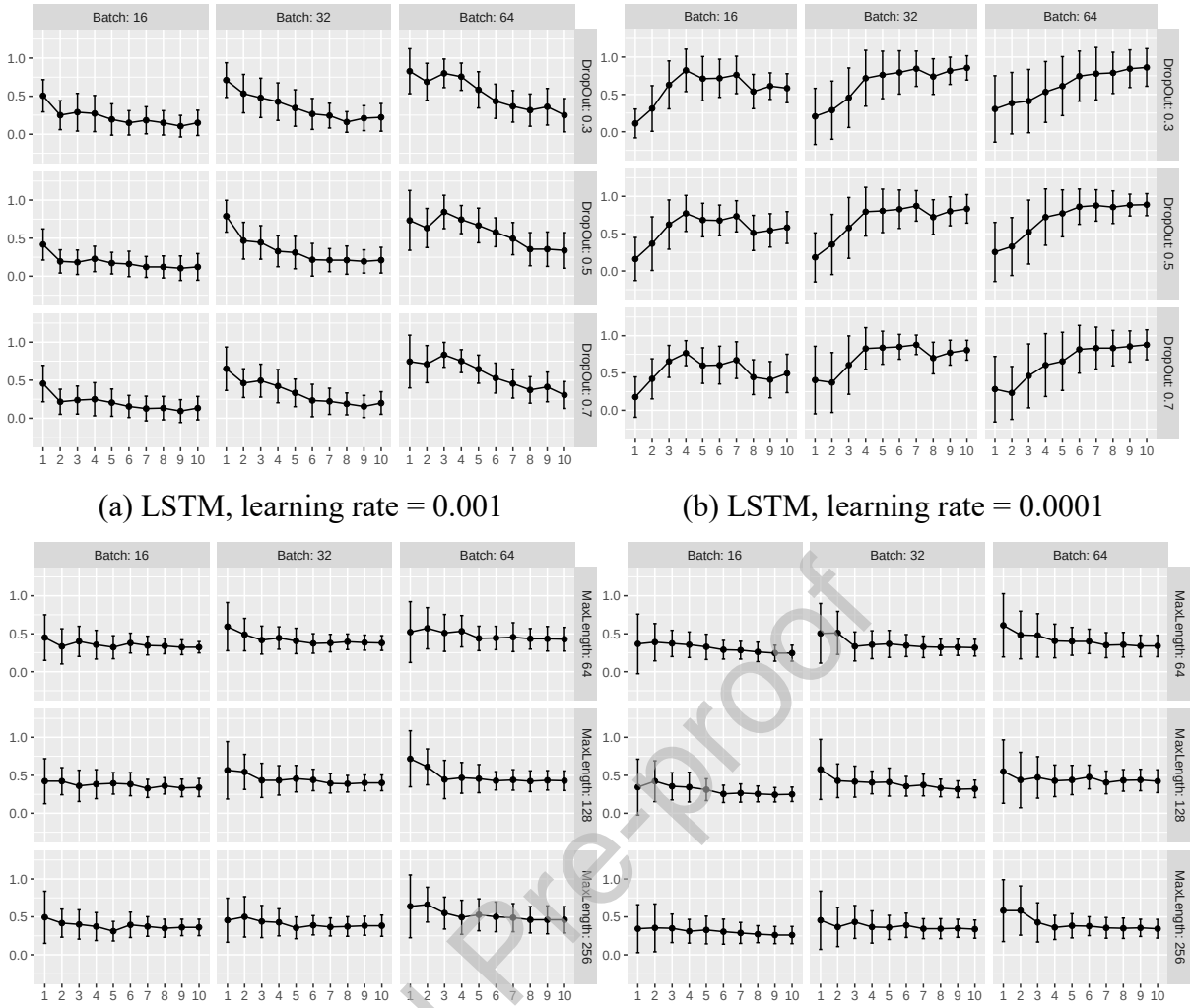
3.2.2 One-argument conditions: $N_{\not CASE}V_{act}$ & $N_{\not CASE}V_{psv}$

The children in Shin (2022a) performed differently by age group and condition. In $N_{\not CASE}V_{act}$, while the five-and-six-year-olds (60%) outperformed the three-and-four-year-olds (42%), the numeric difference in the agent-first response rates was statistically insignificant, indicating that the two groups did not differ considerably in this condition. In contrast, in $N_{\not CASE}V_{psv}$, the rate of agent-first response for the three-and-four-year-olds (59%) increased when compared to this group's performance in its active-voice counterpart, whereas the rate significantly decreased for the five-and-six-year-olds (33%) compared to this group's performance in the active-voice counterpart. These findings indicate that the five-and-six-year-olds reliably interpreted the case-less noun in $N_{\not CASE}V_{psv}$ as the undergoer of an action, suppressing active-voice knowledge in competition, when the computational burden was relaxed. The adult controls demonstrated a very low rate of agent-first response in this condition (10%), indicating their strong theme-first interpretation.

Figures 14 and 15 present the classification performance (coded as Agent-First = 1) of the models per epoch in each condition. Overall, the two architectures demonstrated similar divergence, as shown in the two-argument case-less conditions as the epoch progressed. In $N_{\not CASE}V_{act}$, the LSTM models with a learning rate of 0.001 achieved below-chance performance, indicating that they classified the test stimuli in this condition as Theme-First most of the time, which would have been expected to occur at chance level if the models faithfully simulated the children's response patterns. The models with a learning rate of 0.0001 achieved above-chance performance, indicating that they classified the test stimuli in this condition as Agent-First most of the time, which again would have been expected to occur at chance level if the models faithfully simulated the children's response patterns. These results are inconsistent with Shin (2022a) and Shin and Mun (2023a). In contrast, all

the GPT-2 models were at-chance or slightly below-chance, independently of hyperparameter types, which aligns with Shin and Mun (2023a) but not with Shin (2022a).

In $N_{\text{CASE}}V_{psv}$, the LSTM models with a learning rate of 0.001 achieved below-chance performance, indicating that they classified the test stimuli in this condition as Theme-First most of the time. The LSTM models with a learning rate of 0.0001 showed varying degrees of performance depending on the batch size and the dropout rate. In contrast, whereas all the GPT-2 models with a learning rate of 0.001 showed at-chance performance, all the GPT-2 models with a learning rate of 0.0001 showed below-chance performance, indicating that they classified the test stimuli in this condition as Theme-First most of the time.



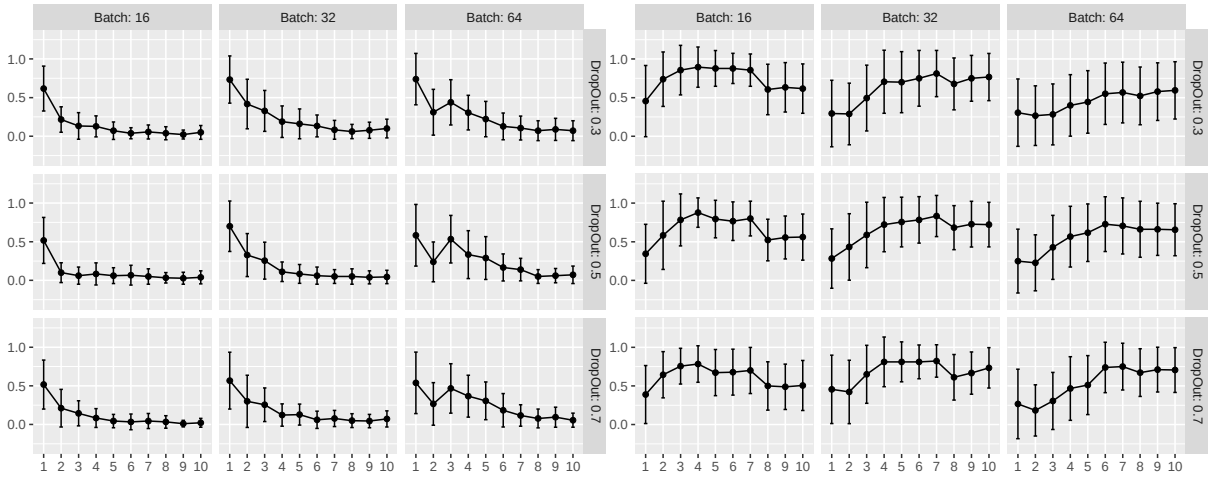
(a) LSTM, learning rate = 0.001

(b) LSTM, learning rate = 0.0001

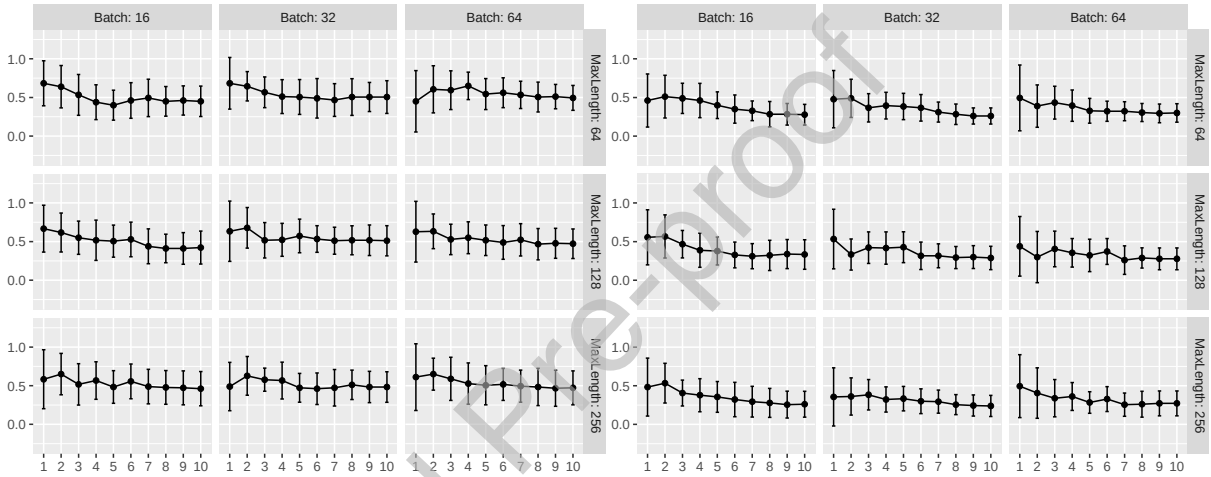
(c) GPT-2, learning rate = 0.001

(d) GPT-2, learning rate = 0.0001

Figure 14. Model performance: $N_{\text{CASE}}V_{act}$. X-axis = epoch; Y-axis = agent-first rate (mean). Error bars = 95% CIs.



(a) Learning rate = 0.001 (b) Learning rate = 0.0001



(a) Learning rate = 0.001 (b) Learning rate = 0.0001

Figure 15. Model performance: $N_{\text{~~CASE~~}}V_{psv}$. X-axis = epoch; Y-axis = agent-first rate (mean). Error bars = 95% CIs.

4. Discussion and Conclusion

4.1 Summary of study

Motivated by the proxy provided by neural networks as biologically inspired models of computation, we developed two neural network models (LSTM; GPT-2) with hyperparameter variations and measured their classification performance on the test sentences used in Shin (2022a) involving scrambling and omission of sentential components to varying degrees. Specifically, we tested if the models could recognise verbal morphology in the suffixal passive construction and conduct the interpretive procedures driven by that morphology (i.e., revision of initial interpretation on the mapping between thematic roles and case markers). We found that, although the performance of these models partially aligned with the children's response patterns observed in Shin (2022a), the models did not faithfully replicate the children's comprehension behaviour pertaining to the suffixal passive. This discrepancy resulted in asymmetries across models, conditions, and hyperparameters.

4.2 Disparity between model performance and child comprehension behaviour

4.2.1 Factors contributing to model performance

The results of this study are attributable to various factors. For instance, whereas Korean caregiver input joins the general characteristics of child-directed speech (Shin, 2022b; cf. Cameron-Faulkner et al., 2003; Snow, 1972; Stoll et al., 2009), it also manifests language-specific properties, such as scrambling and omission of sentential components (see Appendix for the constructional-pattern-wise variability in this respect). Along with the general features of caregiver input, the models may have been sensitive to the specific word order and the type of case markers present in a stimulus during the classification task, particularly as shown in the two-argument case-marked conditions manifesting non-canonical thematic-role ordering ($N_{ACC}N_{NOM}V_{act}$; $N_{NOM}N_{DAT}V_{psv}$) and the case-less conditions ($N_{\cancel{CASE}}N_{\cancel{CASE}}V_{act}$;

$N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$; $N_{\cancel{CASE}}V_{act}$; $N_{\cancel{CASE}}V_{psv}$). This finding aligns with previous reports on language-specific challenges to the automatic processing of Korean (Kim et al., 2007; Shin, 2022b), also partially aligning with Ambridge et al. (2020) showing the failure of modelling human judgements in K'iche'.

Regarding language-specific and construction-specific properties, the models' ability to recognise passive morphology and perform the necessary revision process related to the suffixal passive did not clearly emerge. In the two case-less passive-voice conditions ($N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$; $N_{\cancel{CASE}}V_{psv}$)—the core conditions testing how the models cope with passive morphology and its related interpretive procedures for classification, not all the sub-models succeeded in classifying the test stimuli as Theme-First as intended ($N_{\cancel{CASE}}N_{\cancel{CASE}}V_{psv}$: LSTM, learning rate = 0.001; GPT-2, learning rate = 0.0001, Batch = 16, MaxLength = 256; $N_{\cancel{CASE}}V_{psv}$: LSTM, learning rate = 0.001; LSTM, learning rate = 0.0001, Batch = 64, Dropout = 0.7; GPT-2, learning rate = 0.0001). Moreover, the classification accuracy of model outputs in the case-marked conditions ($N_{NOM}N_{DAT}V_{psv}$; $N_{DAT}N_{NOM}V_{psv}$; $N_{NOM}V_{psv}$; $N_{DAT}V_{psv}$) did not seem to reasonably approximate the children's picture-selection patterns found in Shin (2022a), also manifesting notable by-architecture and by-hyperparameter asymmetries. The precise locus of these asymmetries appears nebulas, as is often the case when interpreting the performance of LLMs in downstream language tasks. However, the disparity between the models' performance and the children's comprehension behaviour in the suffixal passive conditions suggest the following interpretation: neural networks struggle to adapt to language-specific linguistic cues that are language specific, or at least, they process linguistic cues differently from the (developing) human processor does so.

Another factor possibly contributing to the models' performance is the simulation environments in this study. We trained each model with all the transitive-event instances in CHILDES, considering how the children in Shin (2022a) attuned their interpretation to

transitive events before being exposed to the stimuli. Despite this treatment, the models' testing environment may not have fully conformed to what the children partially experienced due to the pre-trained models, mostly comprising adult language features, when constructing each model. Moreover, the test items in the simulations involved no overt acoustic-masking effects (see Table 4) as used in Shin (2022a) that informed the children of something that was somehow hidden (see Table 1). This absence of auditory signals related to the marker(s), which was inevitable due to the simulation settings in which models exclusively processed the textual data, may have unexpectedly affected model performance (cf. Stoyneshka et al., 2010). Taken together, while our simulations were conducted to align with the experimental settings in Shin (2022a) as closely as possible, they stood on somewhat different grounds than the experiments, as is common in modelling research. This difference could have contributed to the observed asymmetry between the models and children, as the models may not have processed the stimuli in the same manner as the children did in the experiments. However, it is important to note that we cannot conclusively attribute this disparity solely to these factors, as these issues remain largely unexplored in the field.

In addition to these factors, algorithmic characteristics of a computational architecture may be a core source of this disparity. Neural networks often utilise contextual information through window-based computation (Haykin, 2009; Kriesel, 2007) when processing data samples. A common practice involves extracting contextual information from formal sequences of words or characters; put differently, neural network models rely heavily on form. While this approach establishes a computational context (cf. Firth, 1957), it differs from the linguistic context encompassing semantic–pragmatic information. Therefore, when models access the meaning or function of a linguistic unit, they resort to the formal co-occurrences in the input rather than drawing directly upon the unit's meaning or function. Moreover, neural networks are designed to generalise existing knowledge (from pre-trained

models and fine-tuning) but are not designed to make reasonable predictions or extrapolations beyond the training space (Marcus, 1998). Deep-learning models attempt to resolve this issue by using massive amounts of data to cover every potential instance of formal co-occurrences; state-of-the-art LLMs with billions of parameters, such as GPT-n, LLaMA, and Bard, benefit from deploying exceedingly—and unrealistically—large training sets. They often yield good performance when handling known inputs but remain unsatisfactory with novel inputs (cf. Choi, 2023), particularly for accessing meaning or function through form (Ettinger et al., 2023; West et al., 2023). More broadly, computational models encounter language usage indirectly and not in a grounded manner; that is, they do not directly engage in language-usage profiles and situations to which language refers (Clark, 1996; McClelland et al., 2020).

Therefore, this algorithmic nature may have caused the models' performance to deviate from the children's response patterns on some test items which could be out of range. The stimuli in Shin (2022a), consisting of animal names as entities, would be new instances for our models in this respect (and also considering the typical composition of transitive sentences in ordinary speech—animate agents and inanimate themes; Dowty, 1991; Langacker, 1991). Some of these stimuli involved scrambling or omission of sentential components, which are also non-typical. These factors may have led the models to malfunction in their operation. The key evidence supporting this argument comes from the models' performance on the conditions in which a simulated learner must determine the thematic role of the first and sole case-less noun only with its presence ($N_{\cancel{CASE}}V_{act}$; $N_{\cancel{CASE}}V_{psv}$) compared to their performance on one-argument case-marked conditions in which a simulated learner has more, and core, information about the first noun's thematic role indicated by a specific case marker next to the noun ($N_{NOM}V_{act}$; $N_{ACC}V_{act}$; $N_{NOM}V_{psv}$; $N_{DAT}V_{psv}$).

Relatedly, the remarkable variations in the models' performance resulting from hyperparameter manipulation further validate our claim regarding the crucial role of algorithmic characteristics in computational models for simulating human language behaviour. Amongst the three hyperparameters chosen for each architecture, we found that the learning rate exerted the greatest influence on adjusting the models' classification behaviour. Given its significance in machine learning (i.e., a hyperparameter that controls the rate at which an algorithm updates or learns parameter values), it likely serves as a proxy for the manner in which humans generalise (linguistic) knowledge. Scholars have debated the process through which learners derive linguistic knowledge from concrete items and apply it to abstract representations—gradual abstraction (conservatism when transferring current knowledge to new items; Ambridge & Lieven, 2015; Goldberg et al., 2004; Theakston et al., 2015) versus early abstraction (rapid generalisation of current knowledge to other relevant items; Fisher, 1996; Gertner et al., 2006; Lidz et al., 2003). If our approach aligns with this concept, the simulations in this study could provide new insights complementing and advancing the literature on how children generalise linguistic knowledge as a function of exposure to linguistic environments and domain-general learning capacities. Nevertheless, we concede that our assertion is based on exploratory observations and is, therefore, speculative. Further examination is warranted.

4.2.2. Factors contributing to child comprehension behaviour

Despite the same pursuit of efficiency in information processing, how a computational model handles language input differs from how the human processor copes with linguistic knowledge. Decades of research have shown that the processor operates to reduce the burden of work currently being executed by immediately mapping form onto function (and vice versa) under simultaneous activation of multiple (non-)linguistic routes, combined with

cognitive-psychological factors (Christianson, 2016; Karimi & Ferreira, 2016; Levy, 2008; McElree, 2000; O'Grady, 2015; Traxler, 2014). In particular, the child processor manifests notable characteristics in its operation due to its developing nature (cf. Omaki & Lidz, 2015), favouring reliable or available cues with a one-to-one mapping relation between form and function (Bates & MacWhinney, 1989; Cameron-Faulkner et al., 2003; Shin, 2021, 2022a; Shin & Mun, 2023b). Given the broad impact of general language-usage experience (Ambridge et al., 2015; Tomasello, 2003), the processor is sensitive to particular linguistic environments in which a target item at hand is situated (Dąbrowska, 2008; Dittmar et al., 2014; Goldberg et al., 2004). The degree to which the current stimulus is informative against the prior language-usage experience also modulates its performance (Dittmar et al., 2008; Shin & Deen, 2023; Stromswold et al., 1985). Furthermore, the contribution of domain-general factors to the processor's operation is sometimes limited or less efficient (Adams & Gathercole, 2000; Diamond, 1985). These aspects collectively modulate how the developing processor adjusts to accomplish sentence comprehension (Choi & Trueswell, 2010; Garcia et al., 2021; Özge et al., 2019; Snedeker & Trueswell, 2004; Suzuki & Kobayashi, 2017).

Reflecting this aspect, the children in Shin (2022a) seemed to make optimal, albeit imperfect or partial, use of the information available at the time, given their learning trajectories. When the children listened to an aural stimulus and were asked to choose one picture that corresponded to the stimulus, they must compute the relative agenthood or themehood between the two arguments with no animacy cue available. Specifically, in the case of the suffixal passive, they must discern verbal morphology indicating the voice and recalibrate the initial, garden-pathed alignments between thematic roles and case markers to formulate a correct interpretation. For this task, the child processor was likely to draw upon multiple morpho-syntactic and semantic cues, including distributional (e.g., mapping between an event representation and a syntactic representation manifested in word order) and local

(e.g., mapping between thematic roles and case markers) ones, which are searchable from their language-usage profiles and are sensitive to usage frequencies. Moreover, their interpretation was likely to be influenced by multiple sources, including event or world knowledge (Friedman, 2000; Snedeker & Trueswell, 2004), memory operation (Kim et al., 2017), task type (Huang et al., 2013), and cognitive bias (e.g., Agent-First strategy; Ambridge et al., 2017; Shin, 2021). This interplay of various (non-)linguistic factors affecting the operation of the child processor may not have been properly captured and modelled by the neural network learners developed in this study.

4.3 Concluding remarks

The present study explored whether and how computational models represent children's language comprehension, focusing on two commonly used neural network architectures in language science, by examining their ability to cope with the Korean suffixal passive construction. Cross-linguistically, acquiring the passive construction is often delayed. In the case of the Korean suffixal passive, given that children have difficulty revising the initial parsing, the interpretive procedures required by passive morphology make acquiring the passive more difficult. Our study revealed that, while computational architectures tested in this study may be able to utilise information about formal co-occurrences to access the intended message to a certain degree, (the outcome of) this process may substantially differ from how a child, as a developing processor, engages in comprehension. This explanation resonates with previous studies showing a notable mismatch between computational models' performance and human-generated data (Chang & Bergen, 2024; Dasgupta et al., 2022; McCoy et al., 2023). We believe that, through its deployment of neural network models with hyperparameter variations and language typologically different from the major languages currently under investigation, our study provides evidence of the limits of the neural

networks' capacity to address child language features. The implications of this study invite subsequent inquiries on the extent to which computational models reveal developmental trajectories of child language that have been unveiled through corpus-based or experimental research. In line with this, comparing model performance across various neural network architectures, manipulating the presence of the patching procedure (i.e., pre-trained-model-only classifiers vs. patched-model classifiers) may provide additional insights into how computational models address child language features.

While this study does not stand on the core assumptions of nativism, such as the poverty of stimulus and innate principles of grammar, our simulations only partially engage with usage-based constructionist approaches that argue for the joint contributions of usage frequency and domain-general learning capacities to shaping learning outcomes, as evidenced in the previous simulation-based studies (Alishahi & Stevenson, 2008; Bannard et al., 2009; Perfors et al., 2011). The current study is limited in computational resources and scope including constructional types, test stimuli, and age range. We thus believe that its implications offer a promising avenue for future studies on this research paradigm in child language development at the intersection of computational methods and techniques.

Notes

1. Abbreviations: ACC = accusative case marker; DAT = dative marker; NOM = nominative case marker; PSV = passive suffix; PST = past tense marker; SE = sentence ender; Strikethrough in grey = obscured; V = verb.
2. Another type of challenge involving passive morphology is that it is morphologically irregular, is unproductive (as they apply only to a limited set of verbs), and overlaps with causative morphology (Sohn, 1999; Yeon, 2015). They were not considered actively in the current study. We hope future research fully reflecting these aspects would replicate the findings of this study.
3. An eojeol refers to a unit with whitespace on both sides that serves as the minimal unit of sentential components. This roughly corresponds to a word in English.
4. See this repository for the code and dataset.
5. One possibility raised was that the caregiver input data may have overridden the adult-language / L1 information during the pre-training stage, akin to catastrophic forgetting observed after fine-tuning a general-purpose model with specific datasets (Kirkpatrick et al., 2017). We acknowledge that our study does not speak to whether this phenomenon occurred in our modelling process, and further research is needed to explore this.

References

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

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

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

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

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

Ambridge, B., & Lieven, E. (2015). A constructivist account of child language acquisition. In B. MacWhinney, & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 478–510). Malden, MA: John Wiley & Sons

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P. M., Bannard, C., Samanta, S., McCauley, S., Arnon, I., Bekman, D., Efrati, A., Berman, R., Narasimhan, B., Sharma, D. M., Nair, R. B., Fukumura, K., Campbell, S., Pye, C., Pixabaj, S. F. C., Paliz, M. M., & Mendoza, M. J. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality

judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. *Cognition, 202*, 104310. https://doi.org/10.1016/j.cognition.2020.104310

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

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

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

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?□. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

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

Borer, H., & Wexler, K. (1987). The maturation of syntax. In T. Roeper, & A. E. Williams (Eds.), *Parameter Setting* (pp. 123–172). Dordrecht: Reidel.

Brooks, P. J., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology, 35*, 29–44. https://doi.org/10.1037/0012-1649.35.1.29

Budzianowski, P., & Vulić, I. (2019). Hello, it's GPT-2 - How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In A. Birch, A. Finch, H. Hayashi, I. Konstas, T. Luong, G. Neubig, Y. Oda, & K. Sudoh (Eds.),

Proceedings of the 3rd Workshop on Neural Generation and Translation (pp. 15–22). Association for Computational Linguistics. https://aclanthology.org/D19-5602

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

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

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

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

Choi, Y. (2023). Common sense: The dark matter of language and intelligence (VLDB 2023 Keynote). *Proceedings of the VLDB Endowment, 16*(12), 4139–4139. https://doi.org/10.14778/3611540.3611638

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

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly Journal of Experimental Psychology, 69*(5), 817–828. https://doi.org/10.1080/17470218.2015.1134603

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

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

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., & Wei, F. (2023). Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023 (pp. 4005–4019). Association for Computational Linguistics. https://aclanthology.org/2023.findings-acl.247

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

Deen, K. U. (2011). The acquisition of the passive. In J. De Villiers, & T. Roeper (Eds.), *Handbook of Generative Approaches to Language Acquisition* (pp. 155–187). Springer Science & Business Media

Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). *Language models show human-like content effects on reasoning*. arXiv preprint arXiv:2207.07051.

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

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

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2014). Familiar verbs are not always easier than novel verbs: How German pre-school children comprehend active

and passive sentences. *Cognitive Science, 38*(1), 128–151. https://doi.org/10.1111/cogs.12066

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547–619. https://doi.org/10.1353/lan.1991.0021

Edwards, C. (2015). Growing pains for deep learning. *Communications of the ACM, 58*(7), 14–16. https://doi.org/10.1145/2771283

Ettinger, A., Hwang, J. D., Pyatkin, V., Bhagavatula, C., & Choi, Y. (2023). "You Are An Expert Linguistic Annotator": Limits of LLMs as analyzers of abstract meaning representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 8250–8263). Association for Computational Linguistics.

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

Friedman, W. J. (2000). The development of children's knowledge of the times of future events. *Child Development, 71*(4), 913–932. https://doi.org/10.1111/1467-8624.00199

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). https://doi.org/10.48550/arXiv.1811.01866

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

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

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

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

Haspelmath, M. (1990). The grammaticization of passive morphology. *Studies in Language, 14*(1), 25–72. https://doi.org/10.1075/sl.14.1.03has

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

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

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

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, 2022-10. https://doi.org/10.1101/2022.10.04.510681

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

Fisher, C. (1996). Structural limits on verb mapping: the role of analogy in children's interpretation of sentences. *Cognitive Psychology, 31*, 41–81. https://doi.org/10.1006/cogp.1996.0012

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

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

Ilharco, G., Wortsman, M., Gadre, S.Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., & Schmidt, L. (2022). Patching open-vocabulary models by interpolating weights. In *the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)* (pp. 29262–29277). https://doi.org/10.48550/arXiv.2208.05592

Jin, K.-S., Kim, M. J., & Song, H.-J. (2015). The development of Korean preschooler' ability to understand transitive sentences using case-markers. *The Korean Journal of Cognitive and Biological Psychology, 28*(3), 75–90.

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

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. https://doi.org/10.1080/17470218.2015.1053951

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, J-B., & Choi, I. (2004). The Korean case system: A unified, constraint-based approach. *Language Research, 40*, 885–921.

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

Kågebäck, M., & Salomonsson, H. (2016). Word sense disambiguation using a bidirectional LSTM. In M. Zock, A. Lenci, & S. Evert (Eds.), *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon* (pp. 51–56). The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/W16-5307/

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kriesel, D. (2007). A brief introduction to neural networks. Available at http://www.dkriesel.com (accessed on 2023-11-07)

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

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

Li, Z., Lyu, K., & Arora, S. (2020). Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, *33*, 14544–14555.

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

Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua, 120*(11), 2546–2556. https://doi.org/10.1016/j.lingua.2010.06.005

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics, 7*, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Ma, Q., Lin, Z., Yan, J., Chen, Z., & Yu, L. (2020). MODE-LSTM: A Parameter-efficient Recurrent Network with Multi-Scale for Sentence Classification. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6705–6715). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.emnlp-main.544/

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

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

Martinez, H. J. V., Heuser, A. L., Yang, C., & Kodner, J. (2023). Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP* (pp. 48–64).

Associations for Computational Linguistics. https://doi.org/10.18653/v1/2023.genbench-1.4

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

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences, 117*(42), 25966–25974. https://doi.org/10.1073/pnas.1910416117

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, *11*, 652–670. https://doi.org/10.1162/tacl_a_00567

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research, 29*(2), 111–123. https://doi.org/10.1023/A:1005184709695

Messenger, K., & Fisher, C. (2018). Mistakes weren't made: Three-year-olds' comprehension of novel-verb passives provides evidence for early abstract syntax. *Cognition, 178*, 118–132. https://doi.org/10.1016/j.cognition.2018.05.002

Moon, S., & Okazaki, N. (2020). Patchbert: Just-in-time, out-of-vocabulary patching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7846–7852). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.631

Ninalga, D. (2023). Cordyceps@ LT-EDI: Patching language-specific homophobia/transphobia classifiers with a multilingual understanding. In *Proceedings*

of the 3rd Workshop on Language Technology for Equality, Diversity and Inclusion (pp. 185–191). https://doi.org/10.48550/arXiv.2309.13561

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

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

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

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

Park, T. (2021). Study on the frequency and causes of the passive in English and Korean in the Gospel of John. *The Journal of Linguistics Science, 98*, 195–213. https://doi.org/10.21296/jls.2021.9.98.195

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & Desmaison, A. (2019). PyTorch: an imperative style, high-performance deep learning library. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc & E. B. Fox (Eds.), *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 8026–8037).

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition, 118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Perkins, L., Feldman, N. H., & Lidz, J. (2022). The power of ignoring: filtering input for argument structure acquisition. *Cognitive Science*, *46*(1), e13080. https://doi.org/10.1111/cogs.13080

Qian, F., Sha, L., Chang, B., Liu, L. c., & Zhang, M. (2017). Syntax Aware LSTM model for Semantic Role Labeling. In K.-W. Chang, M.-W. Chang, V. Srikumar, & A. M. Rush (Eds.), *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing* (pp. 27–32). Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-4305/

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

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition*, *35*(8), 2019–2032. https://doi.org/10.3758/BF03192934

Sagae, K. (2021). Tracking child language development with neural network language models. *Frontiers in Psychology*, *12*, 674402. https://doi.org/10.3389/fpsyg.2021.674402

Schipke, C. S., Knoll, L. J., Friederici, A. D., & Oberecker, R. (2012). Preschool children's interpretation of object-initial sentences: Neural correlates of their behavioral performance. *Developmental Science, 15*, 762–774. https://doi.org/10.1111/j.1467-7687.2012.01167.x

Senior, A., Heigold, G., Ranzato, M. A., & Yang, K. (2013). An empirical study of learning rates in deep neural networks for speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6724–6728). IEEE. https://doi.org/10.1109/ICASSP.2013.6638963

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

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

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

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

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

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

Siewierska, A. (2013). Passive Constructions. In M. S. Dryer & M. Haspelmath (eds.), *WALS Online (v2020.3)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Accessed at http://wals.info/chapter/107 on 2023-10-27.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing.

Cognitive Psychology, 49(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development, 43*(2), 549–565. https://doi.org/10.2307/1127555

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

Stromswold, K., Pinker, S., & Kaplan, R. (1985). Cues for understanding the passive voice. *Papers and Reports on Child Language Development, 24*, 123–130.

Stoll, S., Abbot-Smith, K., & Lieven, E. (2009). Lexically restricted utterances in Russian, German, and English child-directed speech. *Cognitive Science, 33*(1), 75–103. https://doi.org/10.1111/j.1551-6709.2008.01004.x

Stoyneshka, I., Fodor, J. D., & Fernández, E.M. (2010). Phoneme restoration methods for investigating prosodic influences on syntactic processing. *Language and Cognitive Processes, 25*(7-9), 1265–1293. https://doi.org/10.1080/01690961003661192

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification?. In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (eds.), Chinese Computational Linguistics: 18th China National Conference, CCL 2019 proceedings (pp. 194–206), Springer. https://doi.org/10.1007/978-3-030-32381-3_16

Suzuki, T., & Kobayashi, T. (2017). Syntactic cues for inferences about causality in language acquisition: Evidence from an argument-drop language. *Language Learning and Development, 13*(1), 24–37. https://doi.org/10.1080/15475441.2016.1193019

Takase, T., Oyama, S., & Kurihara, M. (2018). Effective neural network training with adaptive learning rate based on training loss. *Neural Networks*, *101*, 68–78. https://doi.org/10.1016/j.neunet.2018.01.016

Theakston, A. L., Ibbotson, P., Freudenthal, D., Lieven, E. V., & Tomasello, M. (2015). Productivity of noun slots in verb frames. *Cognitive Science, 39*(6), 1369–1395. https://doi.org/10.1111/cogs.12216

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences, 18*(11), 605–611. https://doi.org/10.1016/j.tics.2014.08.001

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition, 73*, 89–134. https://doi.org/10.1016/S0010-0277(99)00032-3

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.

de Vries, W., & Nissim, M. (2021). As good as new. How to successfully recycle English GPT-2 to make models for other languages. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 836–846). Association for Computational Linguistics. https://aclanthology.org/2021.findings-acl.74

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

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

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., Jiang, L., Fisher, J., Ravichander, A., Chandu, K., & Newman, B. (2023). The Generative AI paradox: "What it can create, it may not understand". arXiv preprint. https://doi.org/10.48550/arXiv.2311.00059

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing: System demonstrations (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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

Wu, Y., Liu, L., Bae, J., Chow, K.H., Iyengar, A., Pu, C., Wei, W., Yu, L. & Zhang, Q. (2019). Demystifying learning rate policies for high accuracy training of deep neural networks. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 1971–1980). IEEE. https://doi.org/10.1109/BigData47090.2019.9006104

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

Yang, J., Zhang, Y., & Liang, S. (2019). Subword Encoding in Lattice LSTM for Chinese Word Segmentation. 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. 2720–2725). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1278

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9370–9393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.521

Yeon, J. (2015). Passives. In L. Brown, & J. Yeon (Eds.), *The Handbook of Korean linguistics* (pp. 116–136). Oxford: John Wiley & Sons.

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. https://doi.org/10.1038/s41598-021-95392-x

Declaration of interests

 The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.